How to use Microsoft Access to extract data from the 2010 Census Summary File 1

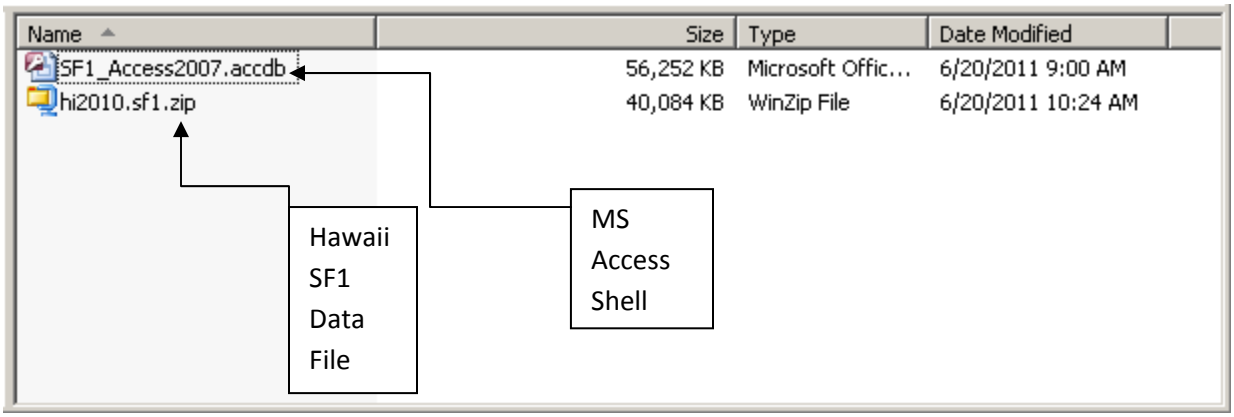
This document provides a step by step example of how to use the Census Bureau provided Microsoft Access database shell to import the FTP version of the 2010 Census Summary File 1 (SF1) data and then extract data from the file.

Background:
The FTP version of the 2010 Census SF1 data is released by state as a series of files within a single compressed .zip file. Each state has forty-eight data segment files and one geo-header file. Each data segment contains one or more data tables. To identify which tables are contained within each segment, please refer to the Table Matrix section in Chapter 6 of the Summary File 1 Technical Documentation (http://www.census.gov/prod/cen2010/doc/sf1.pdf). In order to accommodate the number of fields contained within some segments, pre-processing of those segments is required. Segments number 6, 7, 8, 10, 11, 12, 15, and 38 require pre-processing using a text editing program. Segment 45 requires pre-processing using spreadsheet software. The text editing software and spreadsheet software must be able to accommodate at least 1,020,000 records (rows/lines) of data. This record requirement is to ensure these instructions apply to all states and is set to accommodate the number of rows contained in the segments from the state with the largest number of records. Instructions on how to pre-process these files is embedded in the instructions below.

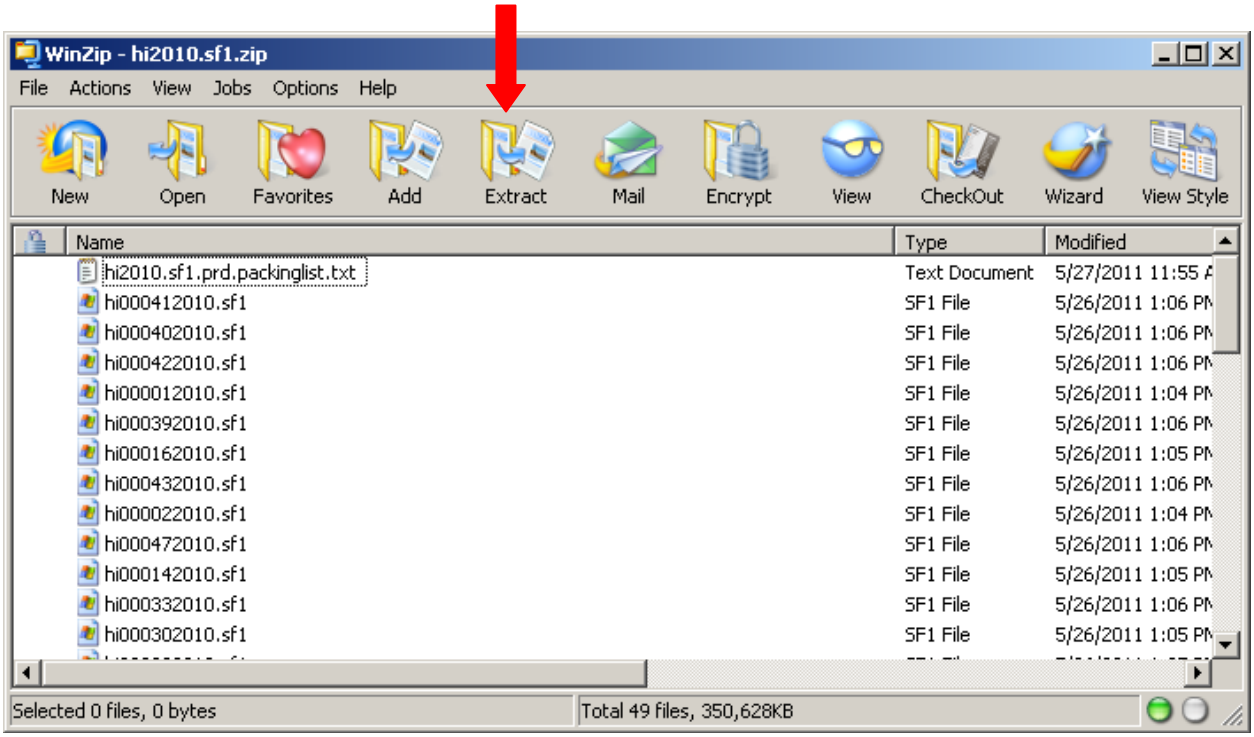
Notes and Assumptions:
1.) MS Access has a 2GB file size limitation. As a result of this limitation, not all segments can be loaded into a single Access database. Multiple Access shells will be required if loading all segments provided as part of a state's SF1 dataset.
2.) The process and steps are the same when using the 2003 or the 2007 version of the shells.
3.) This example uses the Hawaii SF1 data for demonstrating the different import methods.
4.) The data extraction example is for pulling block group summary level data. To identify the summary level number for other geographies, please refer to the technical documentation, Chapter 4 (http://www.census.gov/prod/cen2010/doc/sf1.pdf).
5.) A short list of commonly used summary levels and their GEOID components is available in the GEOID Construction for Matching document. (http://www.census.gov/rdo/pdf/0GEOID_Construction_for_Matching.pdf)
6.) The software used in the examples for the pre-processing of segments are WordPad and Microsoft Excel. Any text editor with a "Find and Replace" function and any spreadsheet software that can accommodate the record limitation listed in the background section can be used. This guidance is not an endorsement of any particular software. It merely provides the structure, using commonly available software, to demonstrate the principles behind preparing, loading, and extracting the data.
7.) Due to the large number of sample tables contained within the shell, users who have their MS Access application set to show hidden objects and system objects may want to change their settings to make them not visible. This does not affect the functionality of the shell.

IMPORTING THE DATA

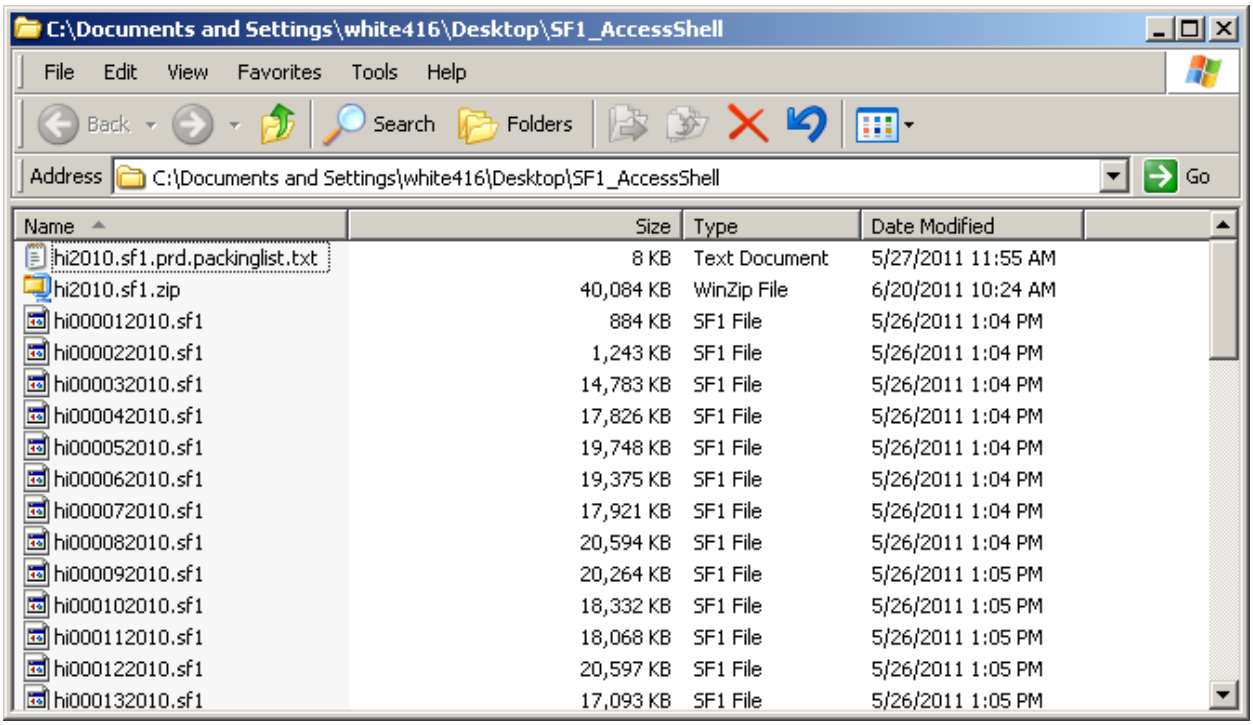
A.) Start by downloading the MS Access shell and the state dataset you want to use. The Access shell and data files can be found at:
http://www2.census.gov/census_2010/04-Summary_File_1/
Once the shell and the zipped data file have been downloaded to your working folder, your working folder will appear similar to this image.



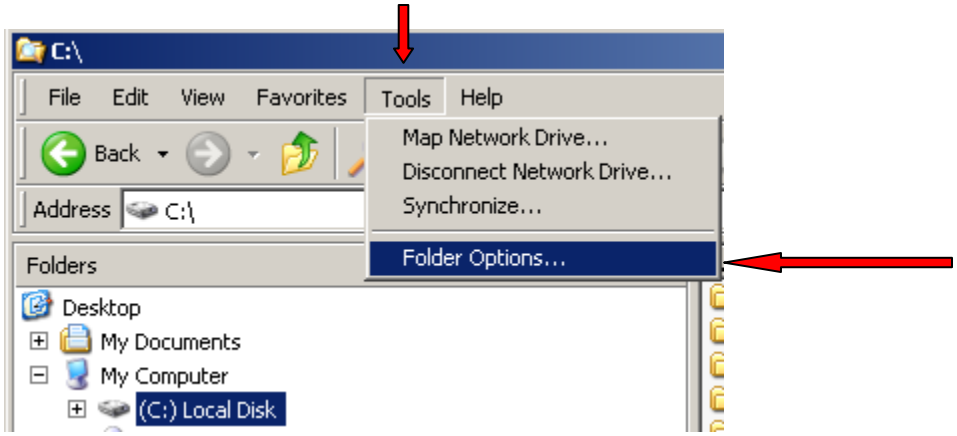
B.) Open the .zip file and extract the files found inside to your working folder. This example uses WinZip. You may use another compression software package or Microsoft Windows to extract these files. The files must be extracted before beginning the import process. This will create uncompressed files in the folder you select during the extraction process.

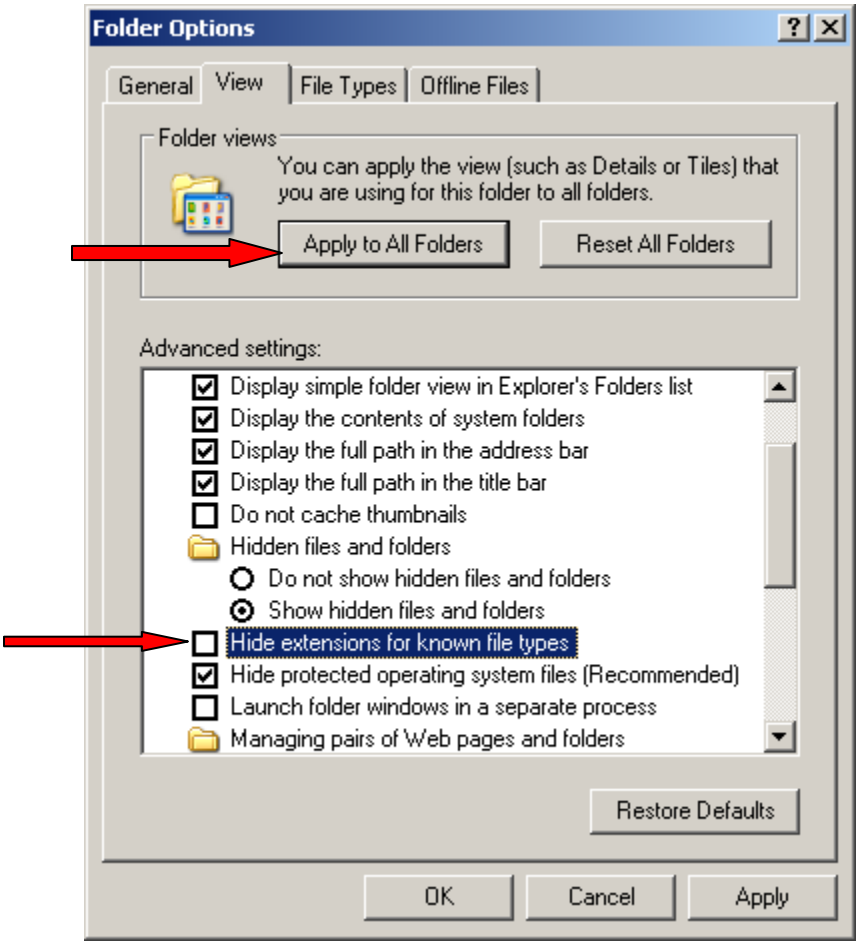


Folder after extraction:

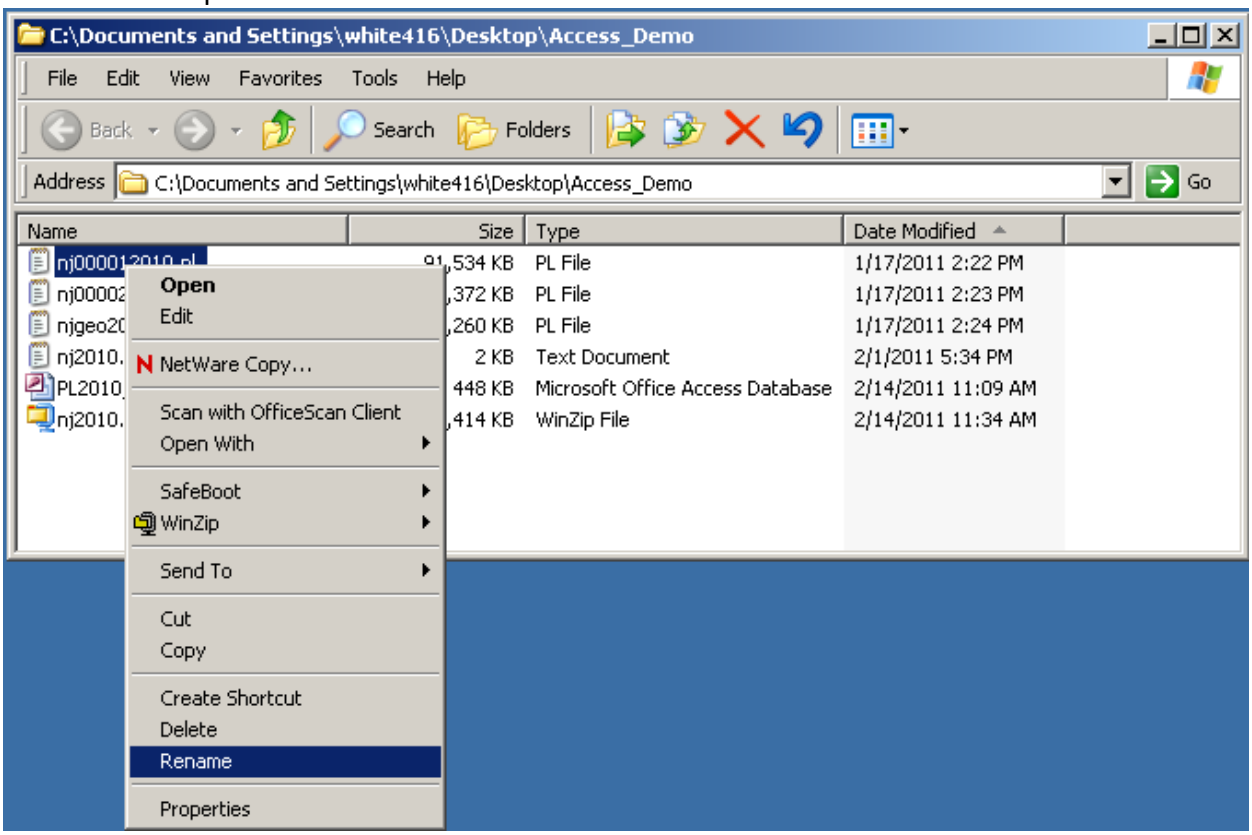


C.) Your Windows environment must be set to show all file extensions. This can be set through the Tools->Folder Options ->View function of Explorer. Make sure the “Hide extensions for known file types” is not checked and then hit Apply to All Folders.

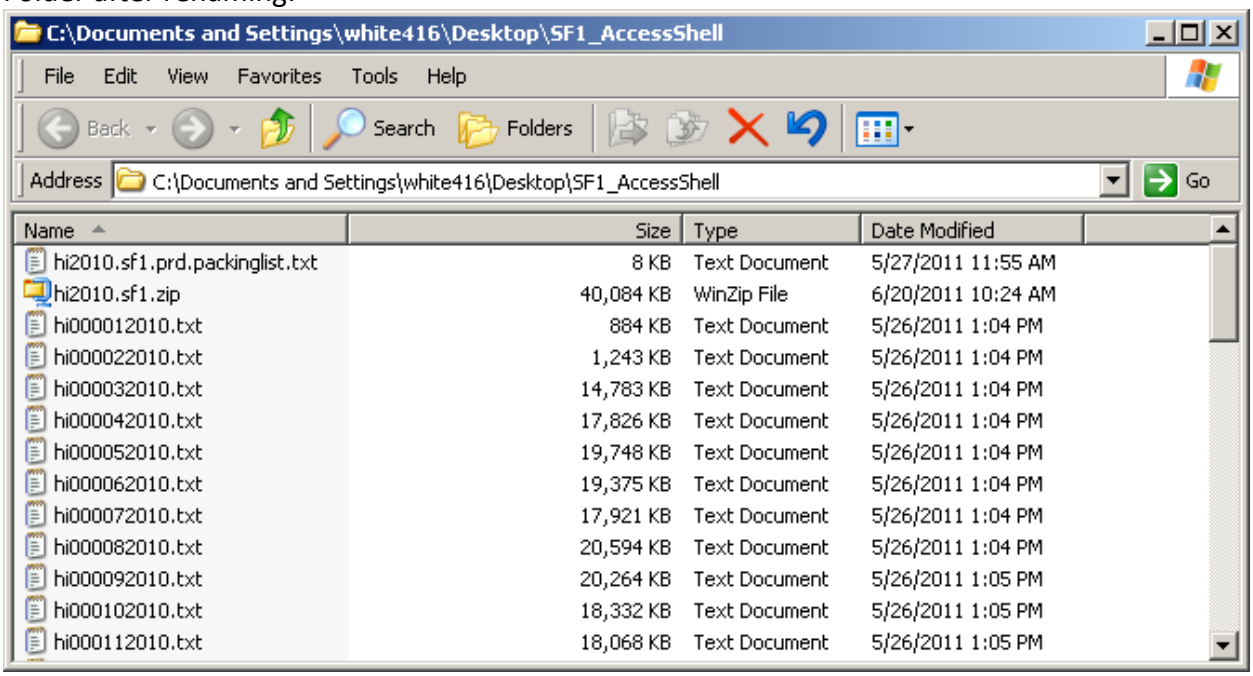




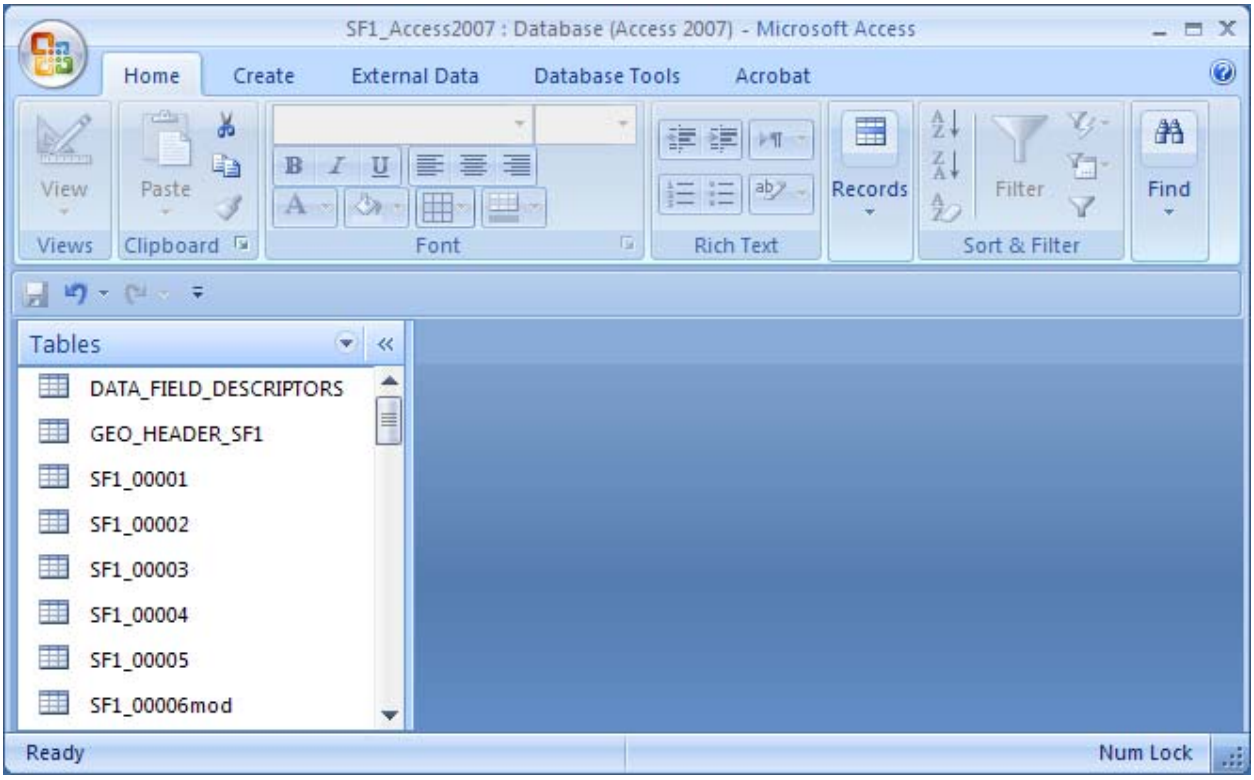
D.) All files with an .sf1 extension must be changed to .txt files. Right click on the first file with a .sf1 extension. Choose “Rename” and change the .sf1 portion of the name to .txt and hit Enter. Repeat for each file with a .sf1 extension.



Folder after renaming:



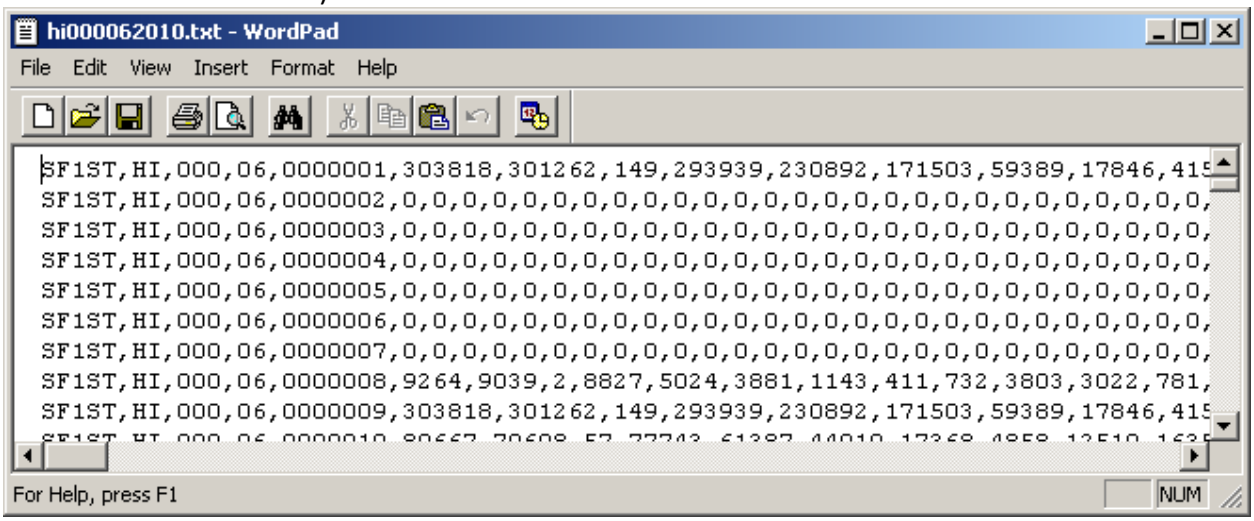
E.) Open the MS Access Shell. You'll see example tables of each segment (SF1_00005, SF1_00006mod, etc) and the geo-header (GEO_HEADER_SF1) along with a table version of the data dictionary (DATA_FIELD_DESCRIPTORS). Note: Those segments listed with "mod" in their name require pre-processing using the Find/Replace method described later in this document. Segments listed as PT1 and PT2 require pre-processing using the spreadsheet method, also described later in this document.



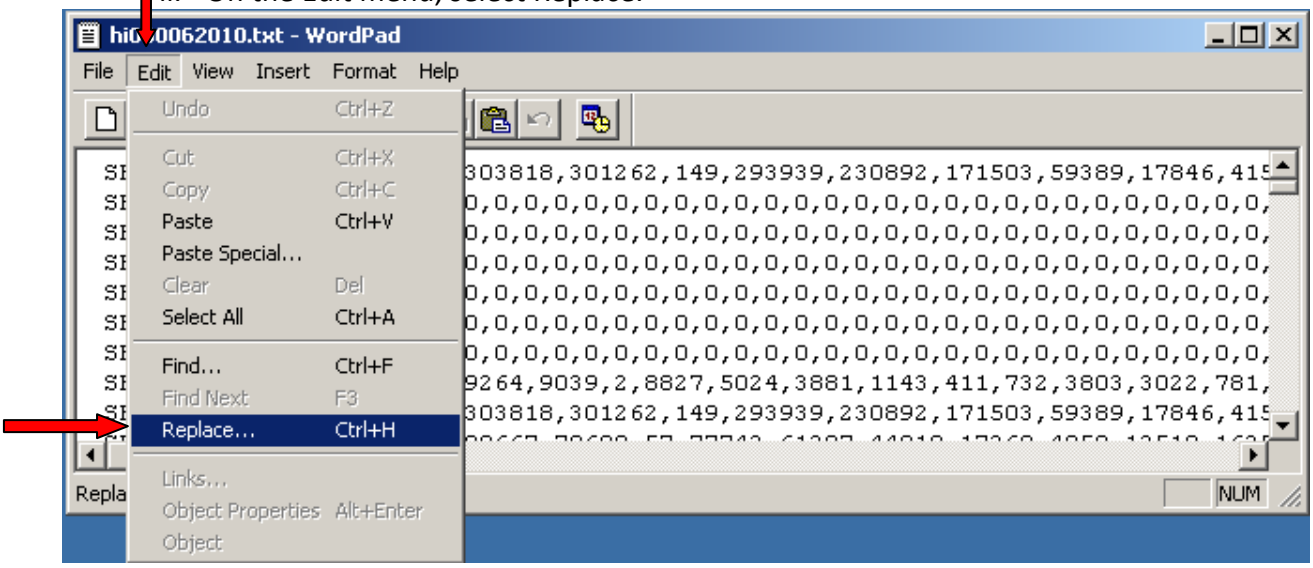
F.) To pre-process segments 6, 7, 8, 10, 11, 12, 15, and 38 begin with subsection “a” of this step (step F). If working with segment 45 go to subsection “b” of this step (step F). For all other segments, proceed to step G.

a. Pre-processing segments 6, 7, 8, 10, 11, 12, 15, or 38 requires the use of text editing software that has a find and replace function. In order to prepare these segments for use with the Access shell it is necessary to “Find” the data contained within the first four fields of the segment and “Replace” them with nothing. The result will be a segment that begins with the field LOGRECNO rather than FILEID. These segments can be readily identified within the Access database by the appending of the text “mod” to both the table example and the import script’s name. **NOTE: Performing this modification can take several minutes for smaller states and several hours for larger states.**

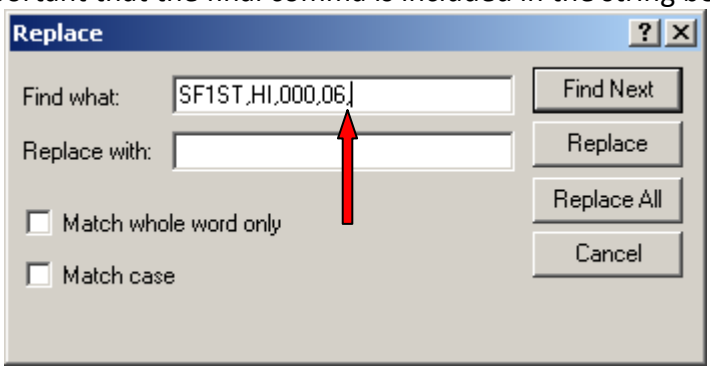
i. Open the segment using your text editor software. This example uses WordPad. (Notepad is not acceptable software for performing this modification.)



ii. On the Edit menu, select Replace.



iii. In the “Find what:” entry box type in “SF1ST,<ST>,000,,” where <ST> is the two letter capitalized postal abbreviation for the state in which you are working and is the two digit segment number for the segment you are preparing. If the segment number is a single digit it should be preceded by a zero. This example shows the entry for Hawaii segment 6. Leave the “Replace with:” field blank and click the Replace All button. This may take some time to run as it is processing the entire segment. NOTE: It is important that the final comma is included in the string being replaced.



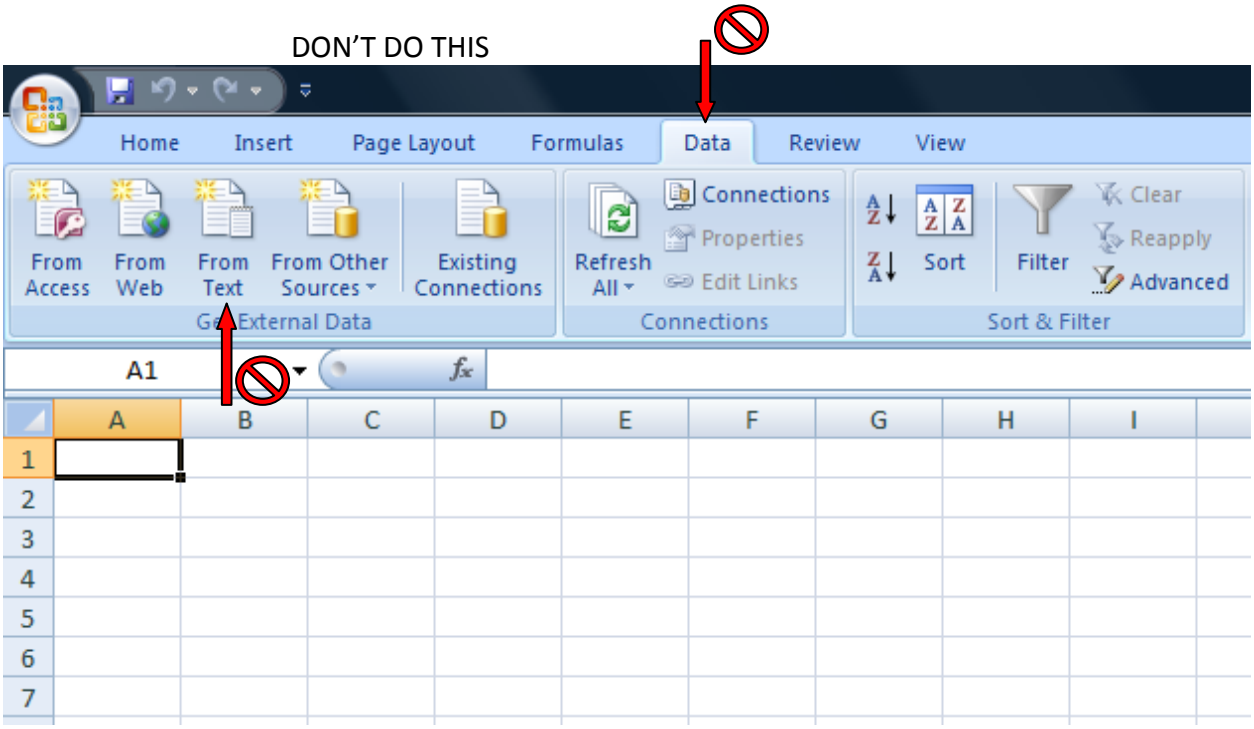
iv. Once the operation is complete, click the Cancel button to close the Find/Replace window and then save the file – a suggested file naming convention is to keep the original name and add “mod” to the end. This will keep it consistent with the shell’s designations.

v. This completes the pre-processing requirement for the segment. Proceed to step G of the Importing Data section to complete the import of the segment into the Access database.

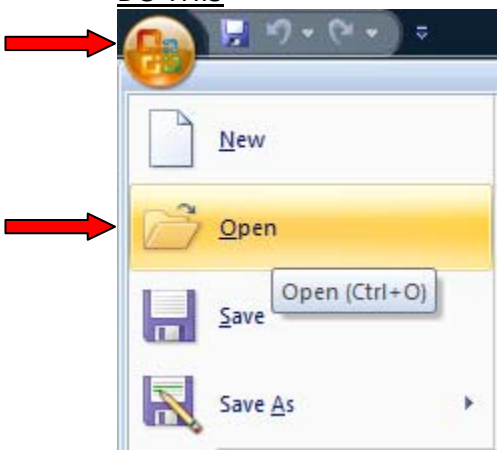
b. Pre-processing segment 45 requires the use of spreadsheet software. In order to prepare this segment for use with the Access shell import scripts it is necessary to split the segment into a Part 1 and a Part 2. The result will be two segments, one containing all of the tables of segment 45 except for table H17C (Part1) and another segment containing only table H17C (Part2). These segments can be readily identified within the Access database by the appending of the text "PT1" and "PT2" to both the table example and the import script's name. The split of the segment must be done precisely as described in order for the file to work with the import scripts provided.

i. Open the segment using your spreadsheet software. This example uses Microsoft Excel 2007. How you open the file is critical. When using Excel, do not use the Data->Import from Text function. This will give you an error and not allow the import of all columns. It is essential that you use the "Office Button" and then select Open.

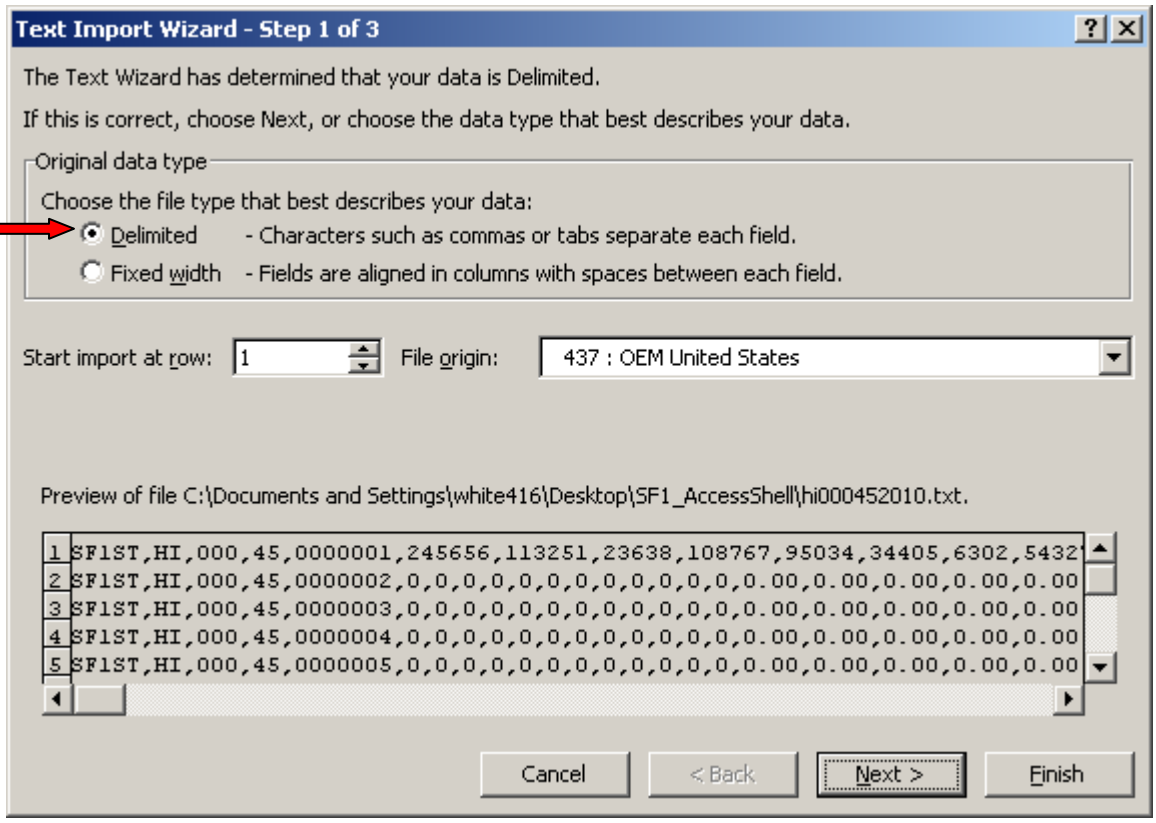
DON'T DO THIS



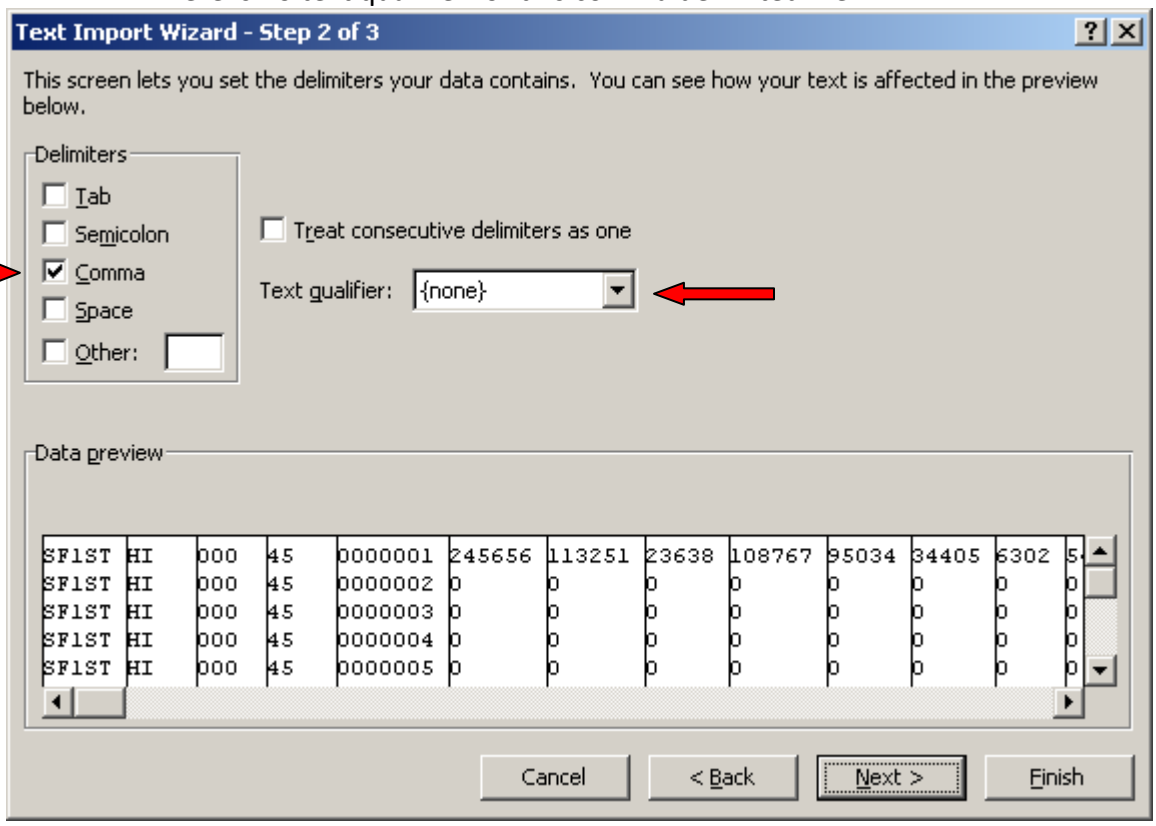
DO THIS



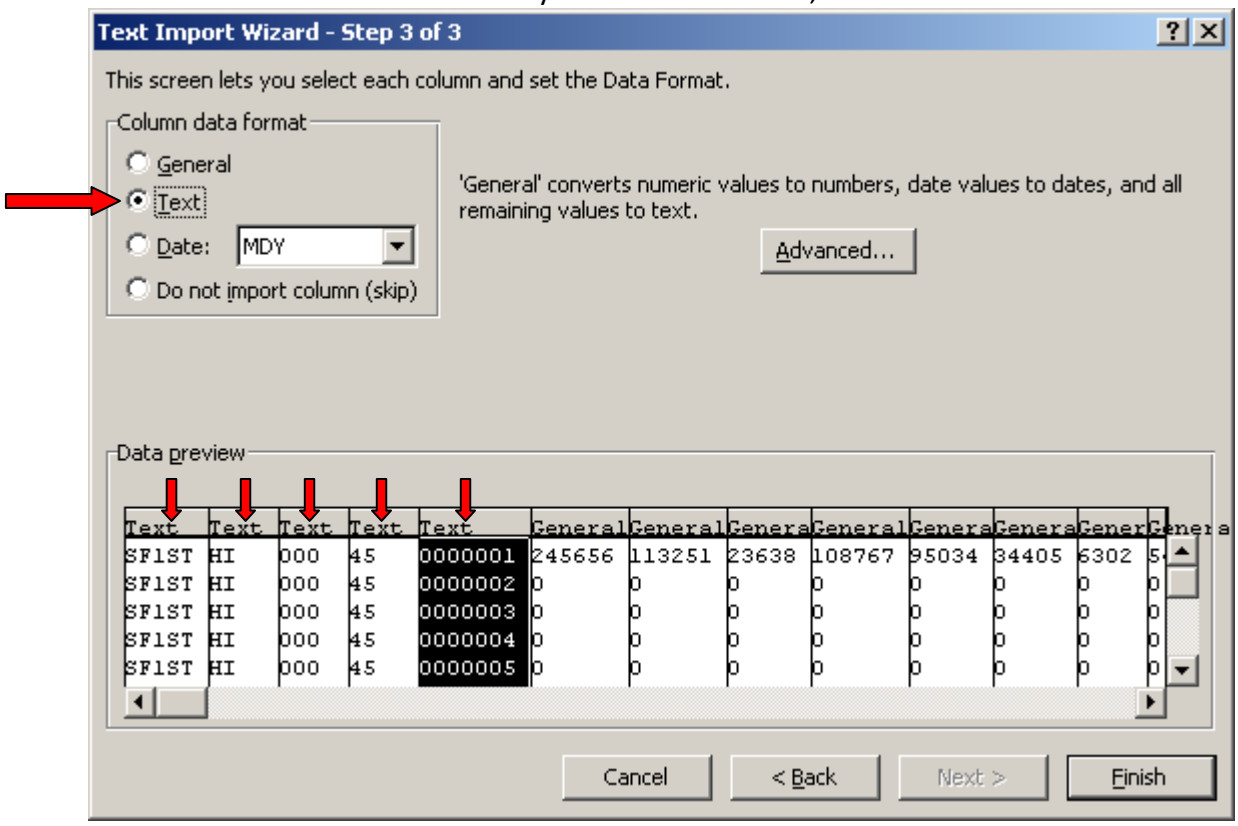
ii. The file is a comma delimited file and the data starts in the first row.



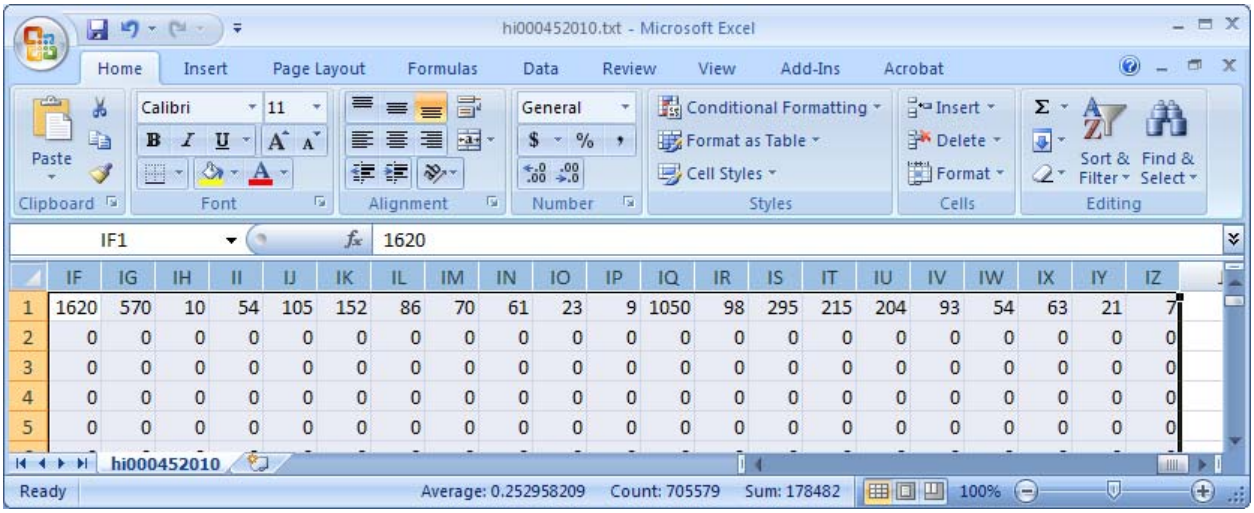
iii. There is no text qualifier for this comma delimited file.

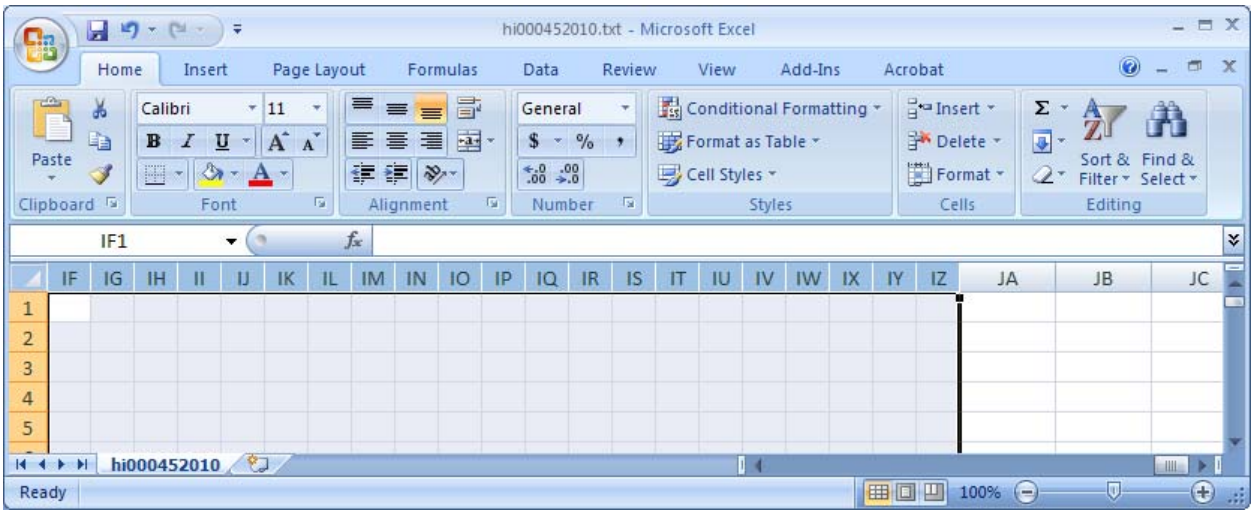


iv. Set the first five columns to “Text” by clicking on their column header and selecting “Text” from the radio buttons. This has to be done for each of the five columns individually. Once these are set, click the Finish button.



v. To create the Part 1 file, identify the last 21 columns in the spreadsheet and delete them from the file. In Excel these will be columns IF to IZ.





vi. Use Save As to save the file as a .csv file. It is recommended that when saving the file you append PT1 to the name to distinguish it from, and not overwrite, the original. Close Excel.

vii. Locate the newly created .csv file and rename it to become a .txt file.

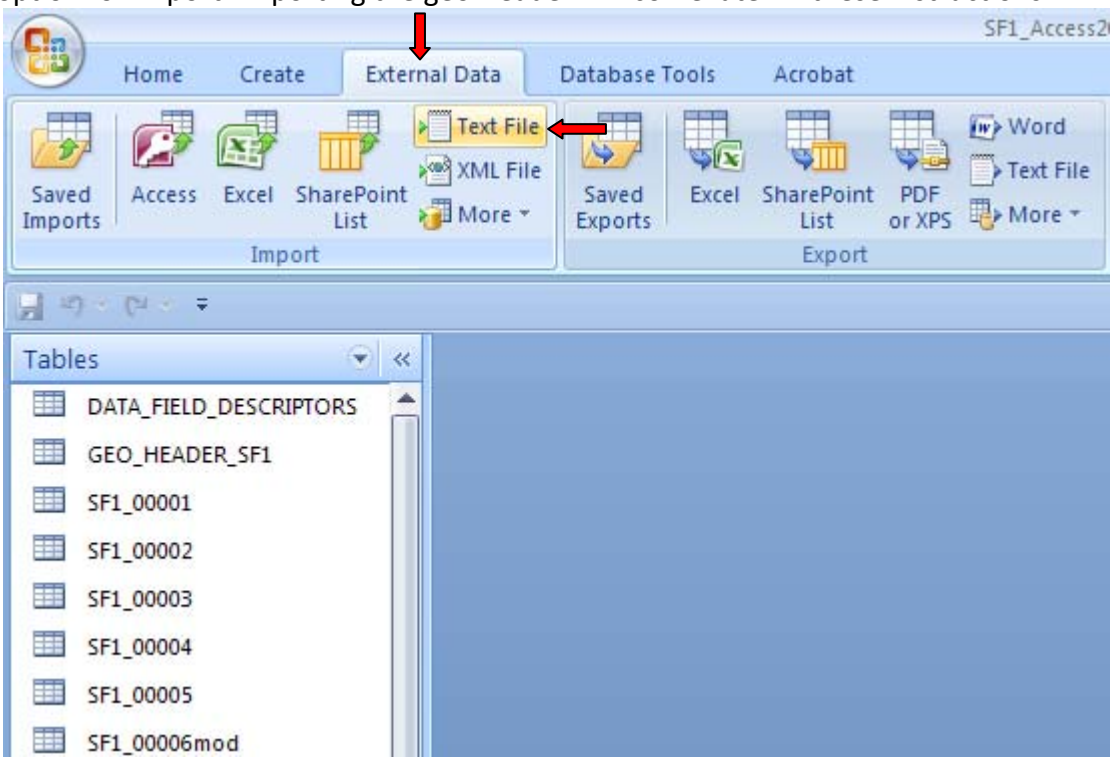
viii. To generate the Part 2 version of the file, open the original file in your spreadsheet software and delete all but the first five and last twenty one columns. In Excel this will be deleting columns F to IE. It is important to delete the columns not just the data in the columns. Your new table should be only 26 adjoining columns. In Excel these become columns A to Z.

ix. Use Save As to save the file as a .csv file. It is recommended that when saving the file you append PT2 to the name to distinguish it from the original. Close Excel

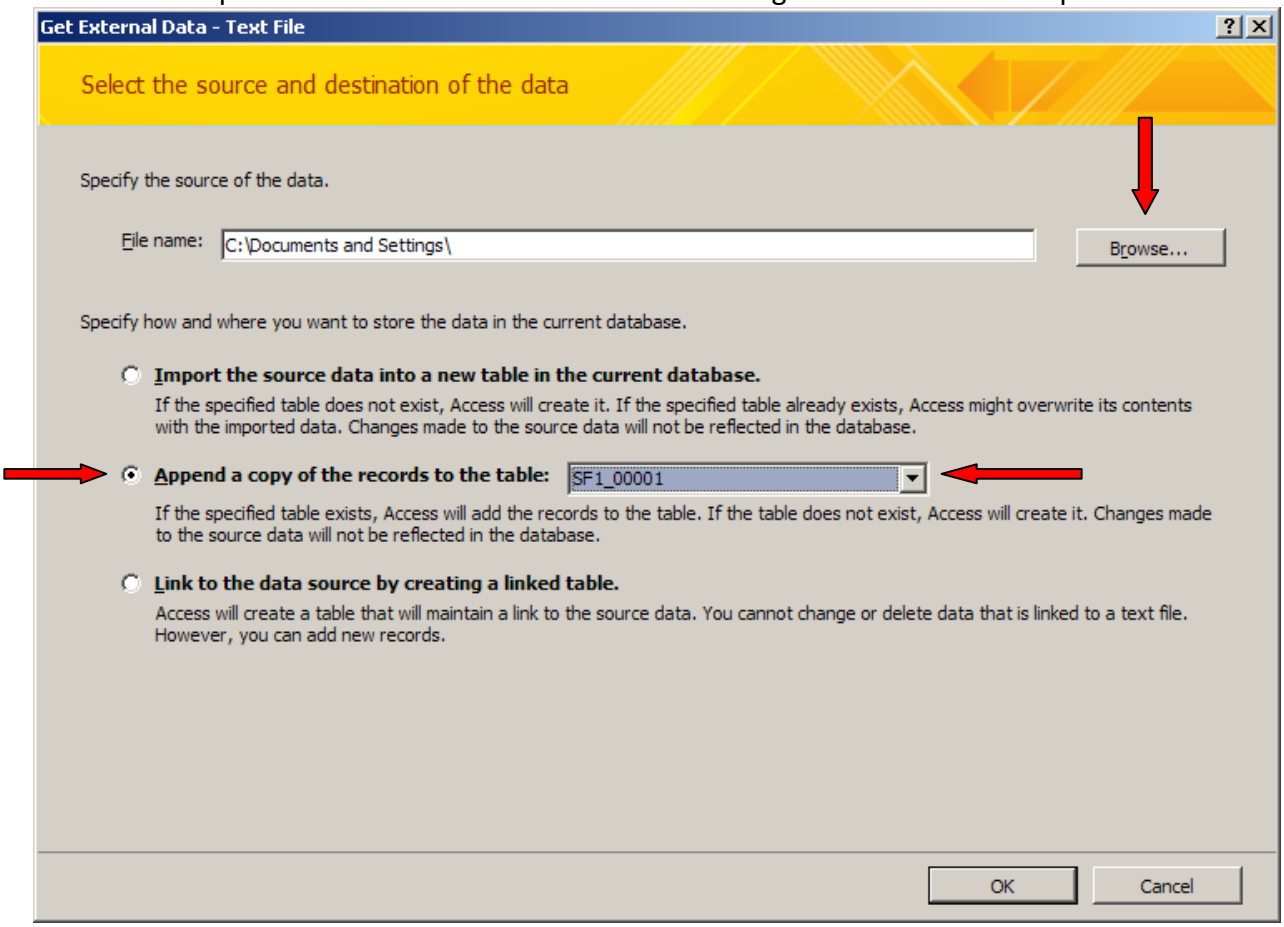
x. Locate the newly created .csv file and rename it to become a .txt file.

xi. Proceed to step G of the Importing Data section to complete the import of these new “PT1 & PT2” segments into the Access database.

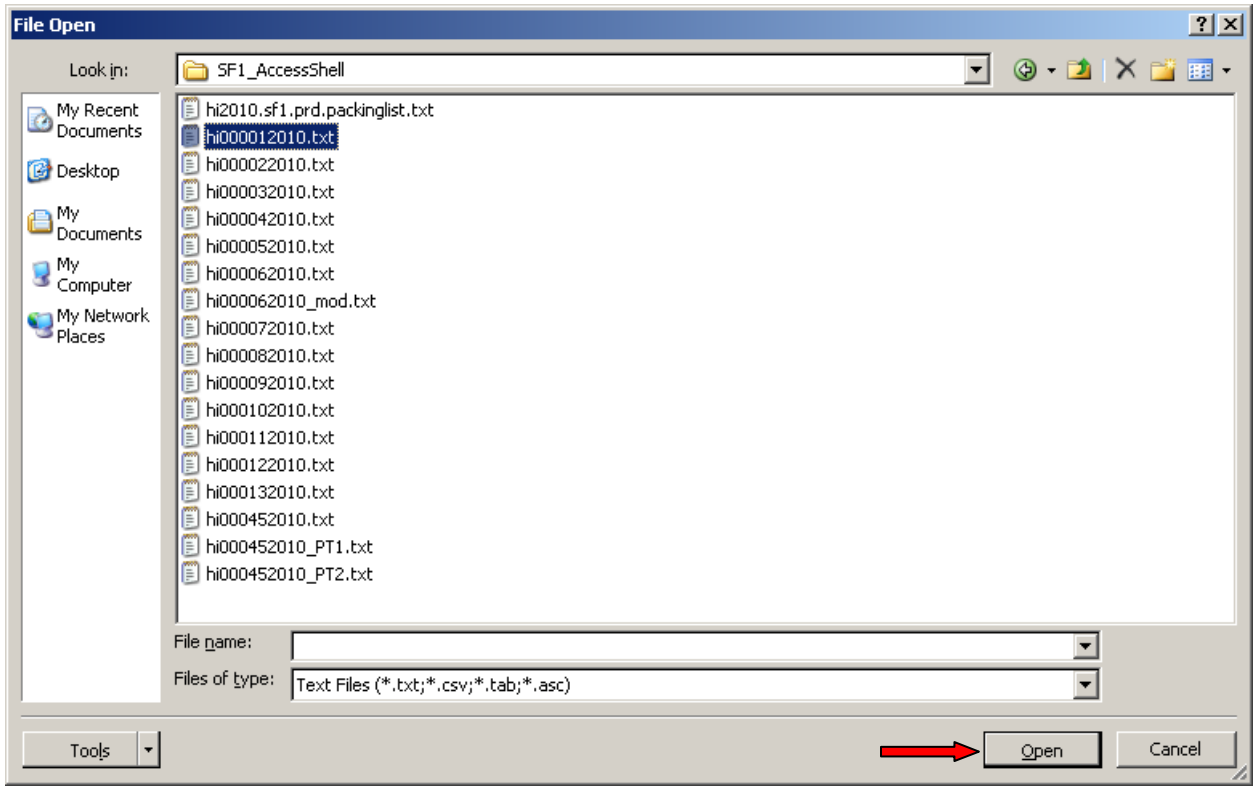
G.) To import a data segment, click on the “External Data” menu and select the “Text File” option for import. Importing the geo-header will come later in these instructions.



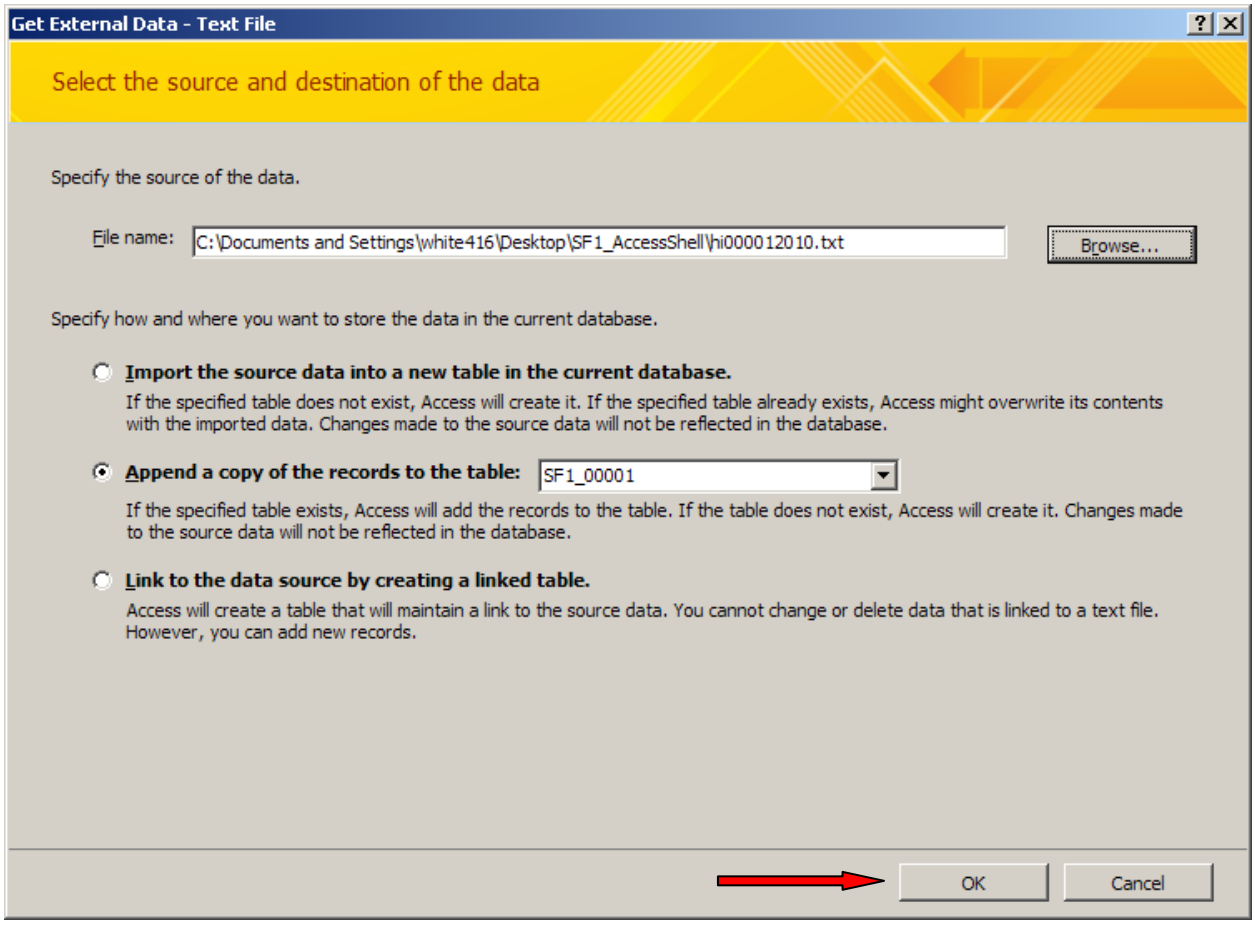
H.) Make sure the radio button selection at the bottom of the import screen is on “Append a copy of the records to the table:” and select the sample segment table you are importing into in the drop down list. Use the browse button to navigate to the files for import.



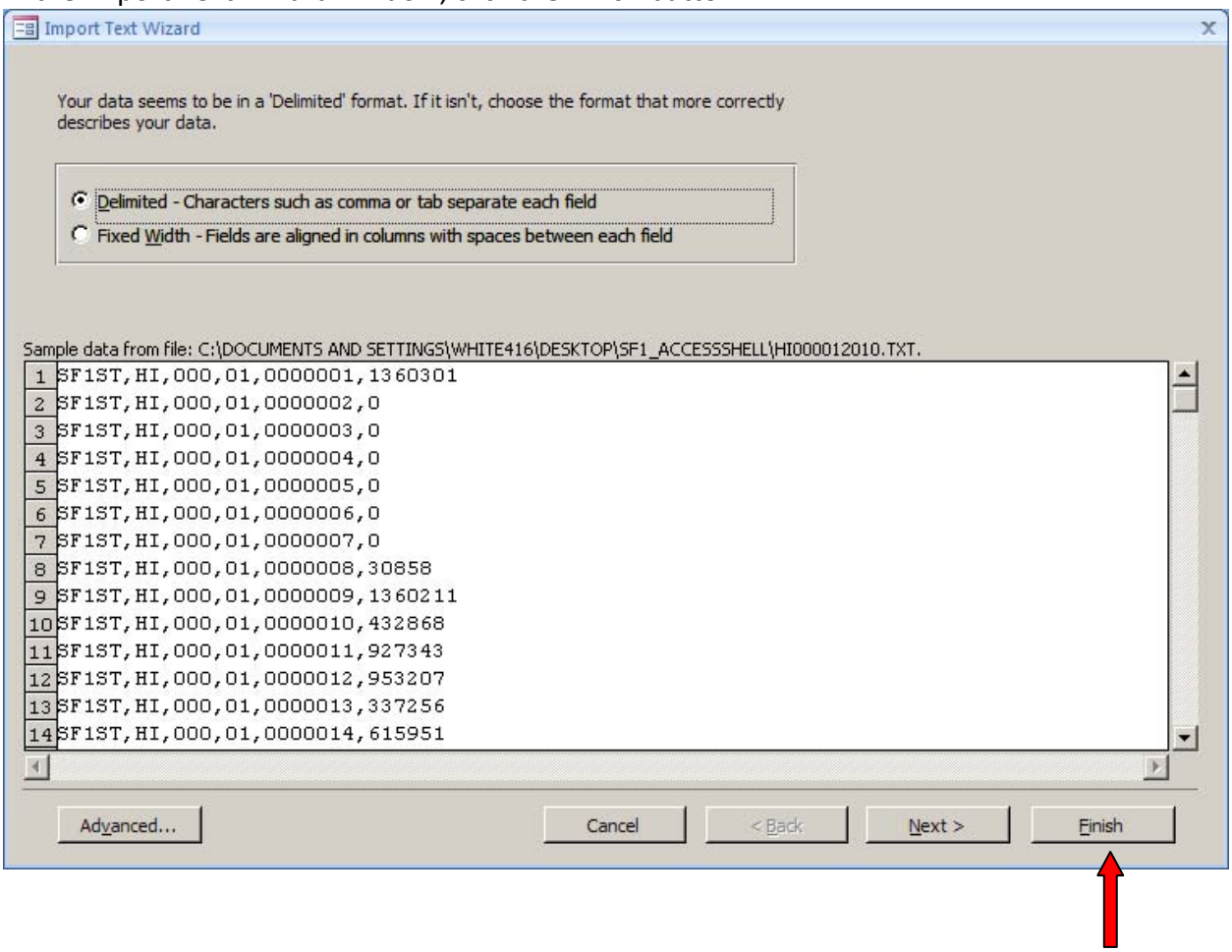
I.) Choose the first file to import. This example imports Hawaii Segment 01. Select the file in the upper window and click the Open button. The segment naming structure for the data segments is <ST>0002010. The geo-header is <ST>geo2010. The <ST> is the two letter postal abbreviation for the state represented by the file and is the two digit segment number. When the segment number is a single digit it is preceded by an additional zero.



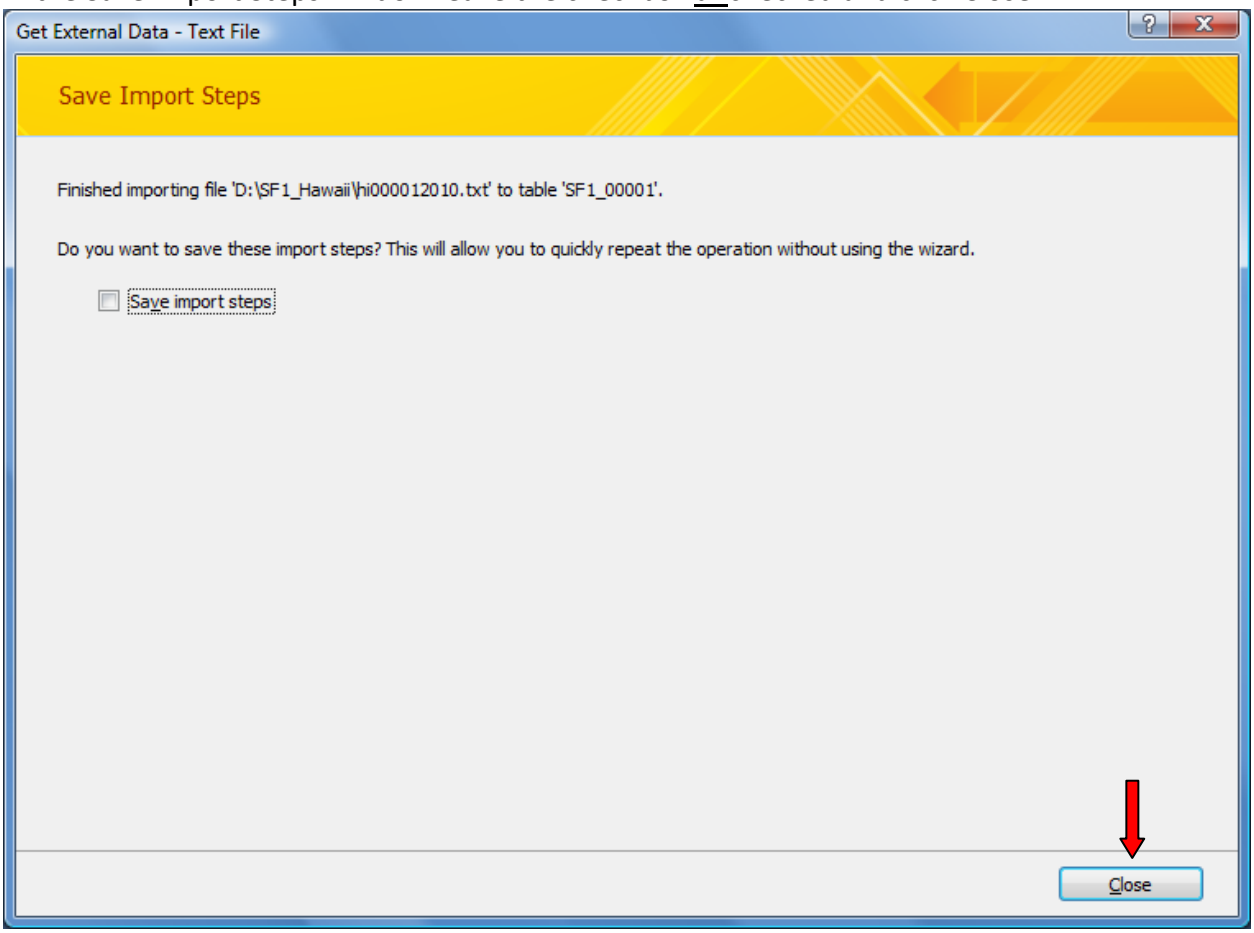
J.) Click the OK button in the Get External Data – Text File window.



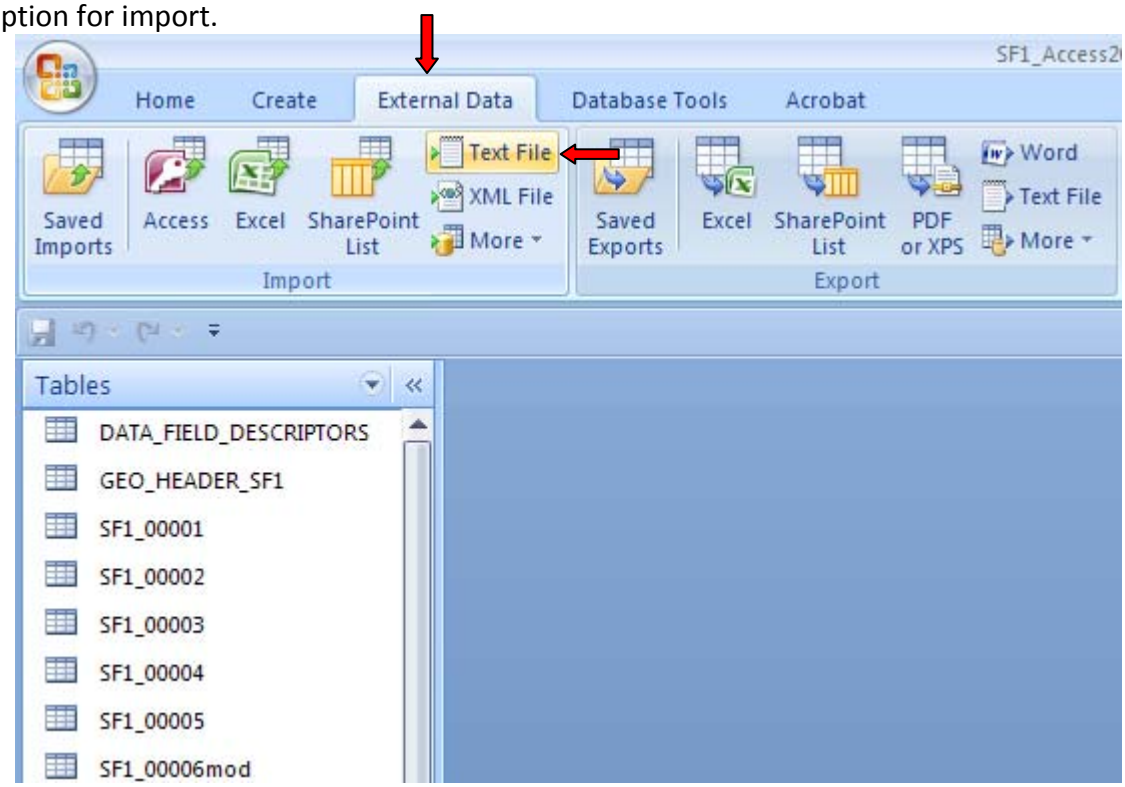
K.) In the Import Text Wizard window, click the Finish button.



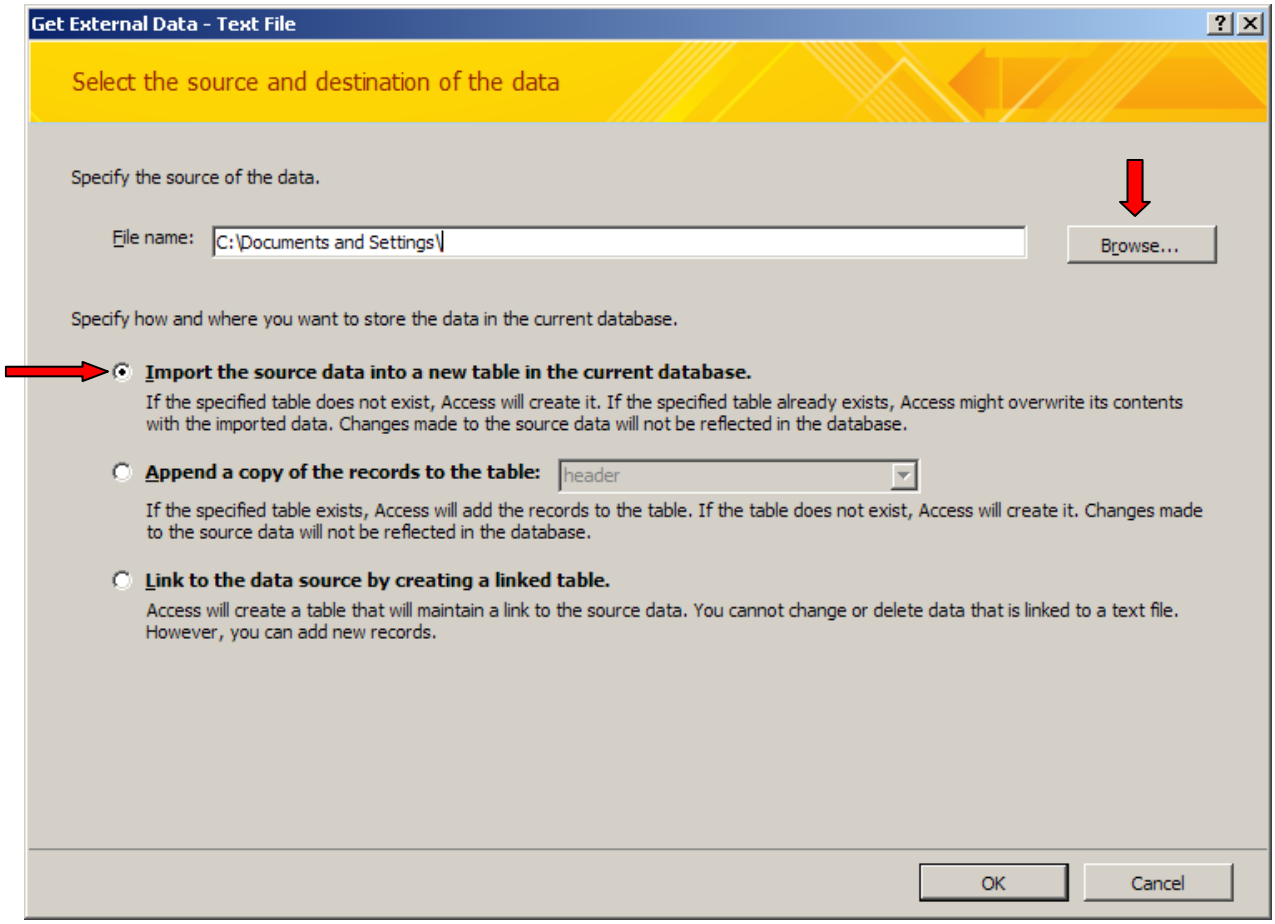
L.) In the Save Import Steps window leave the checkbox unchecked and click Close.



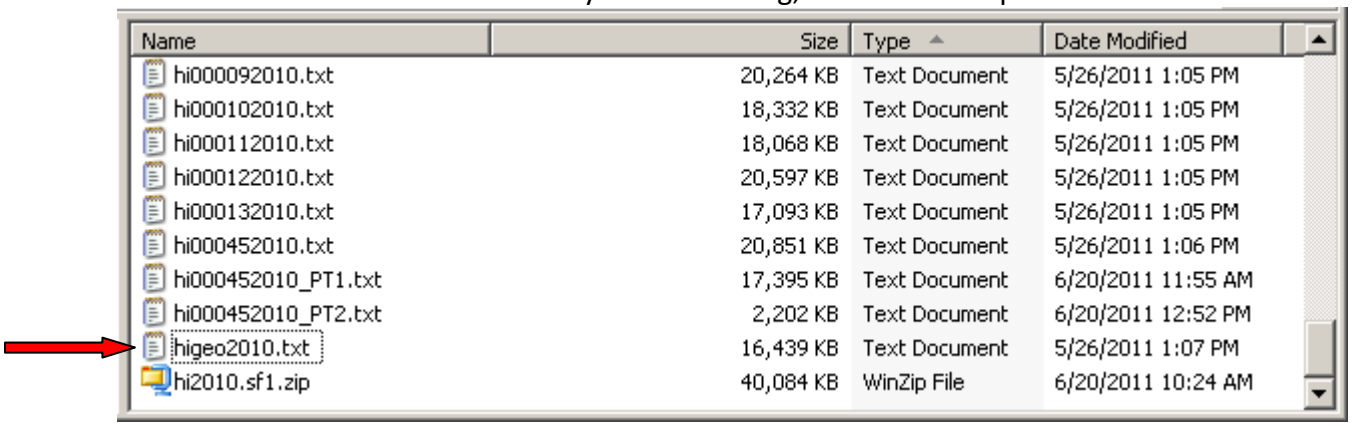
M.) To import the geo-header file, click on the "External Data" menu and select the "Text File" option for import.



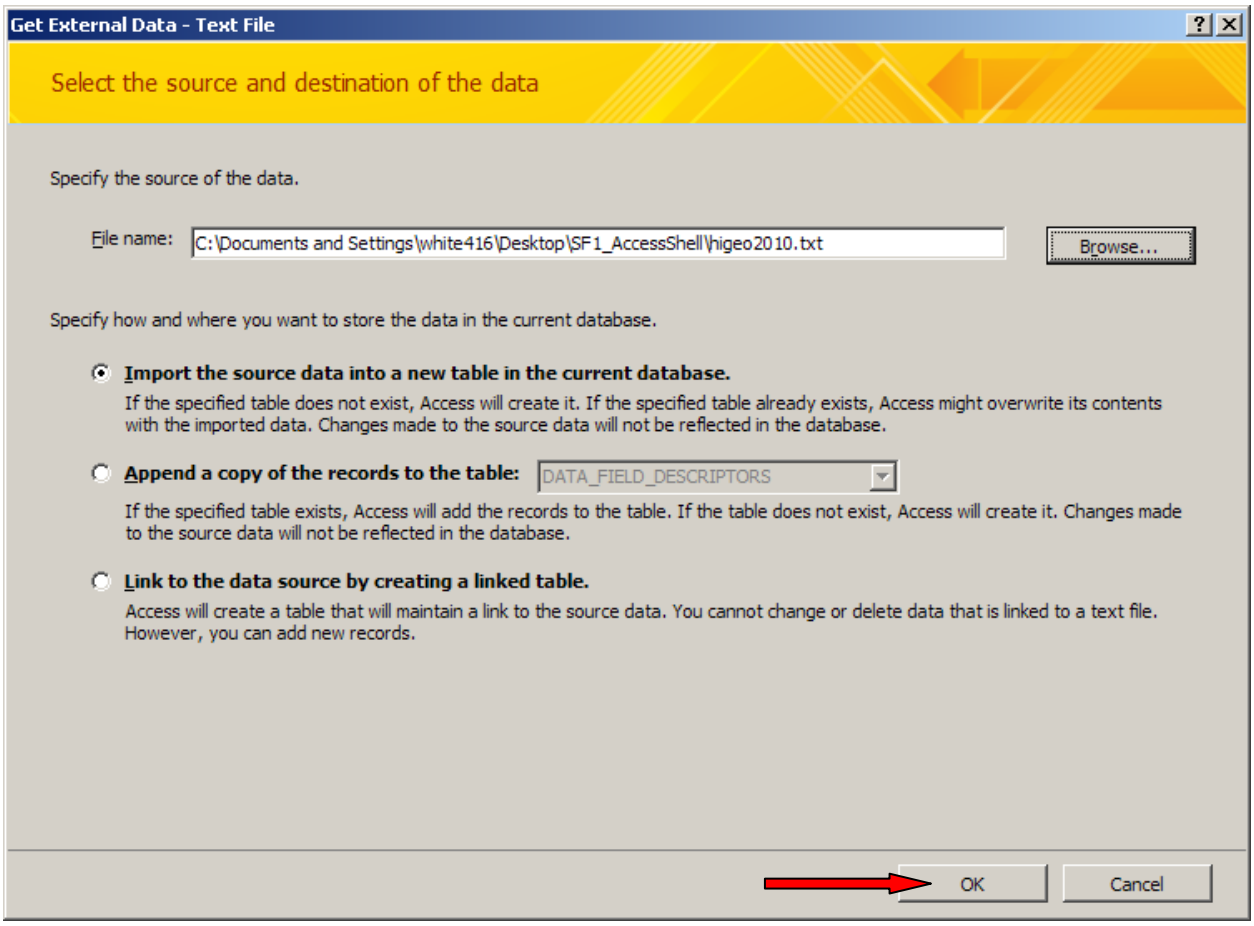
N.) Make sure the radio button selection at the bottom of the import screen is on “Import the source data into a new table in the current database.” Use the browse button to navigate to the file for import.



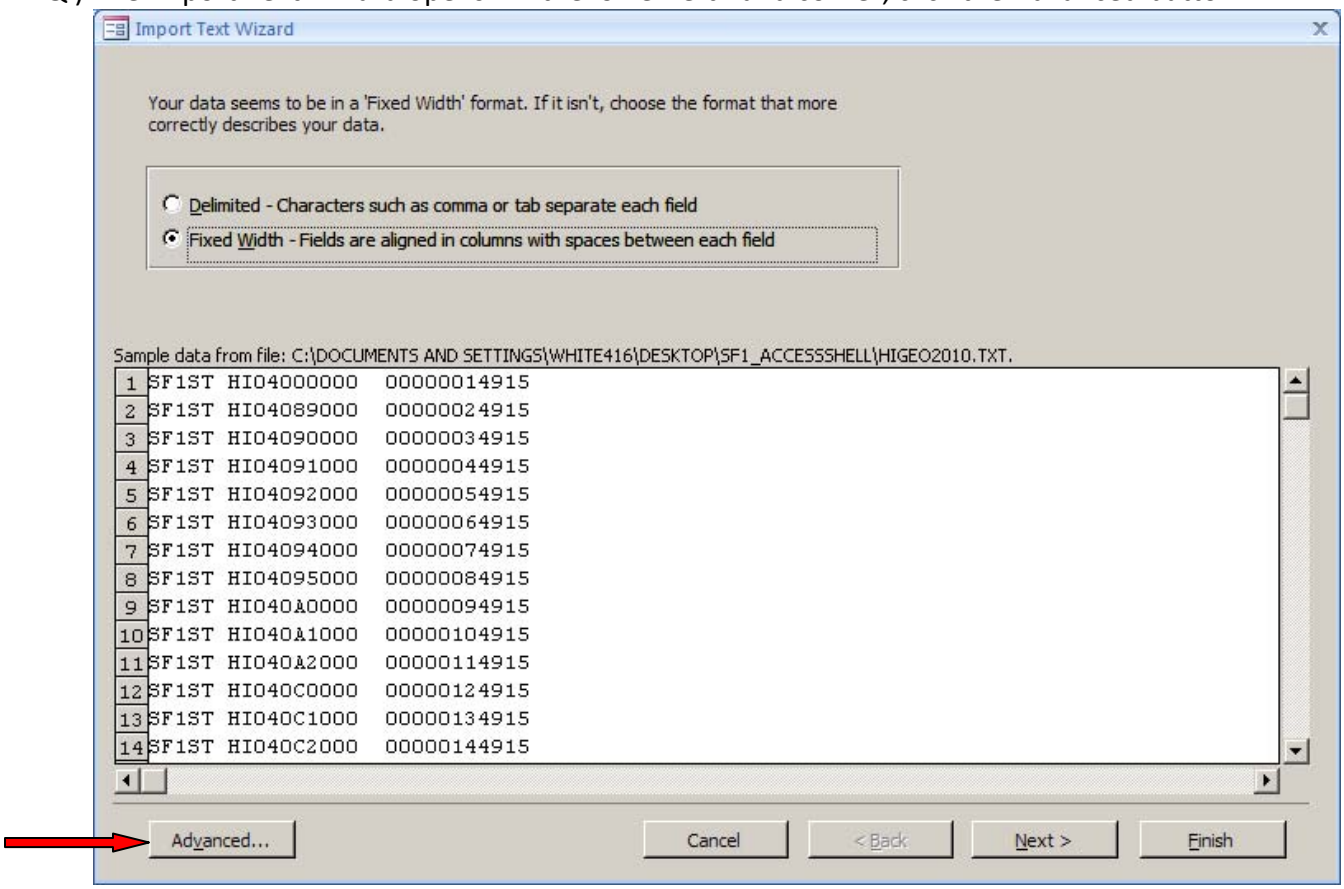
O.) Select the <ST>geo2010.txt in the upper window, where <ST> is the two letter postal abbreviation for the state with which you are working, and click the Open button.



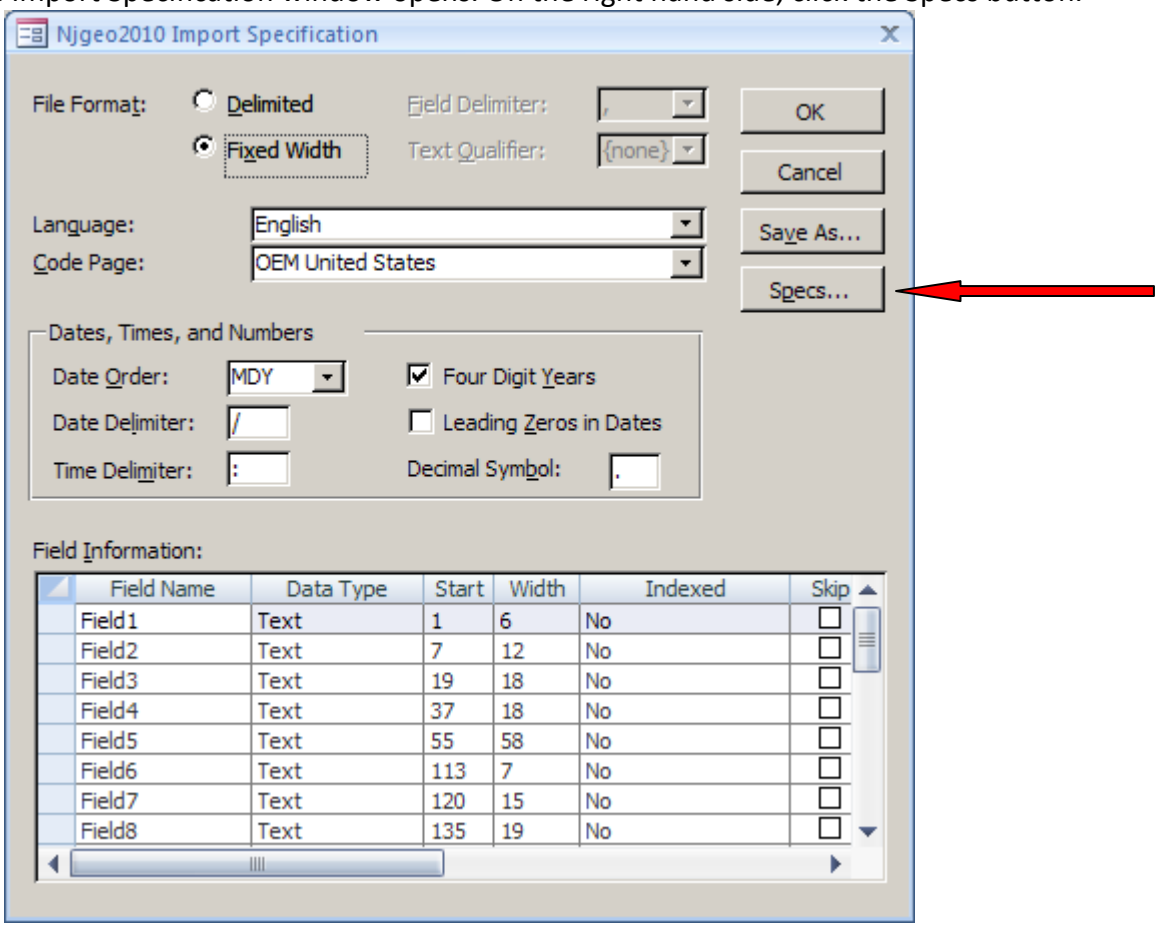
P.) Click the OK button in the Get External Data – Text File window.



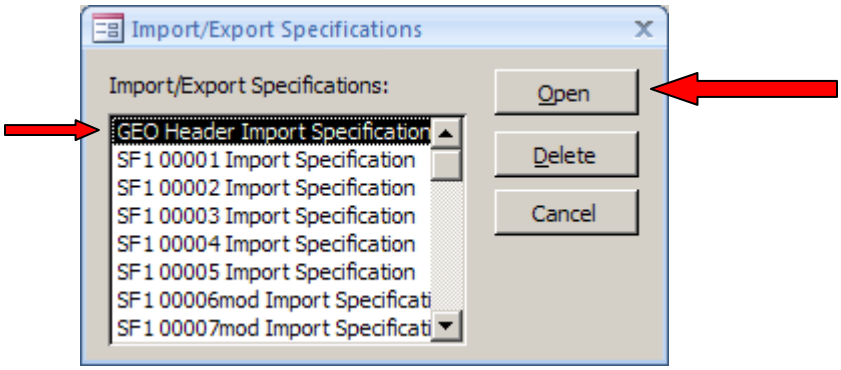
Q.) The Import Text Wizard opens. In the lower left hand corner, click the Advanced button.



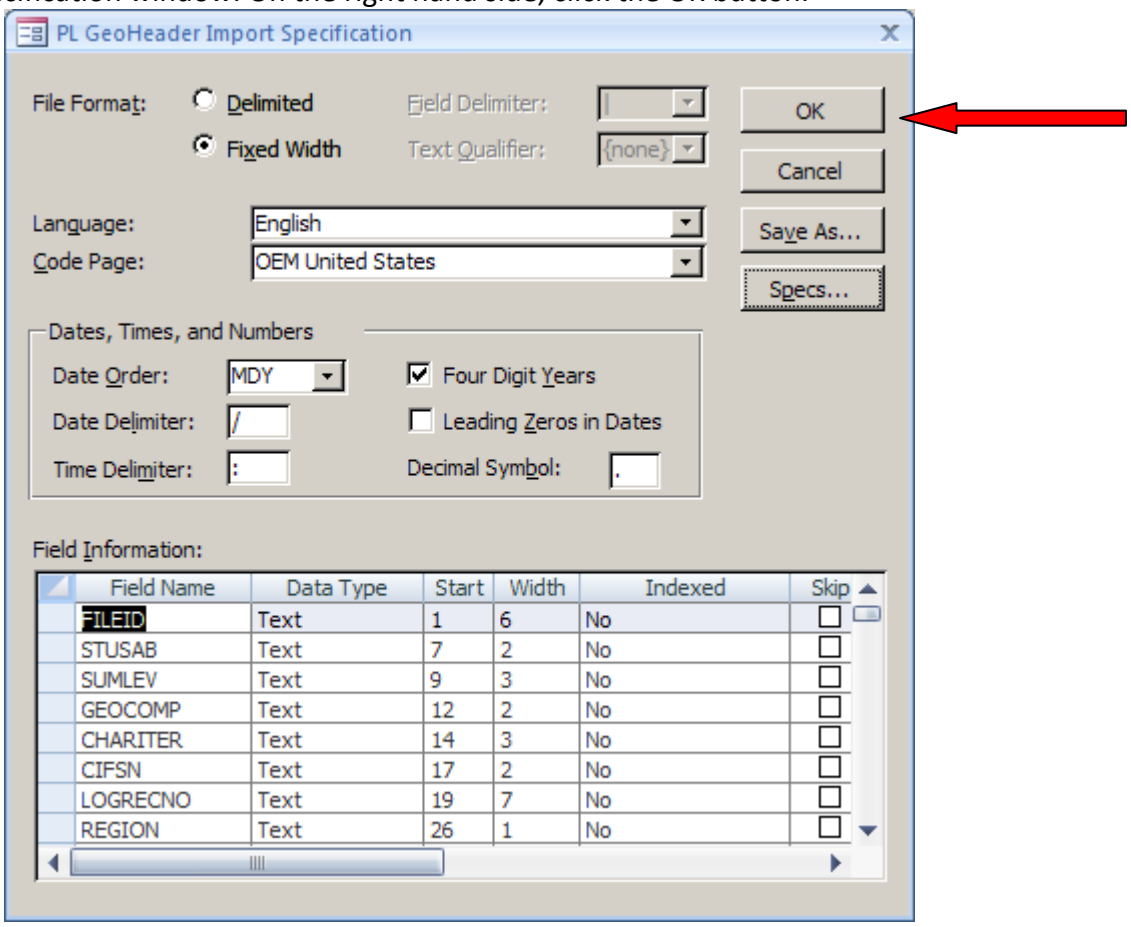
R.) The Import Specification window opens. On the right hand side, click the Specs button.



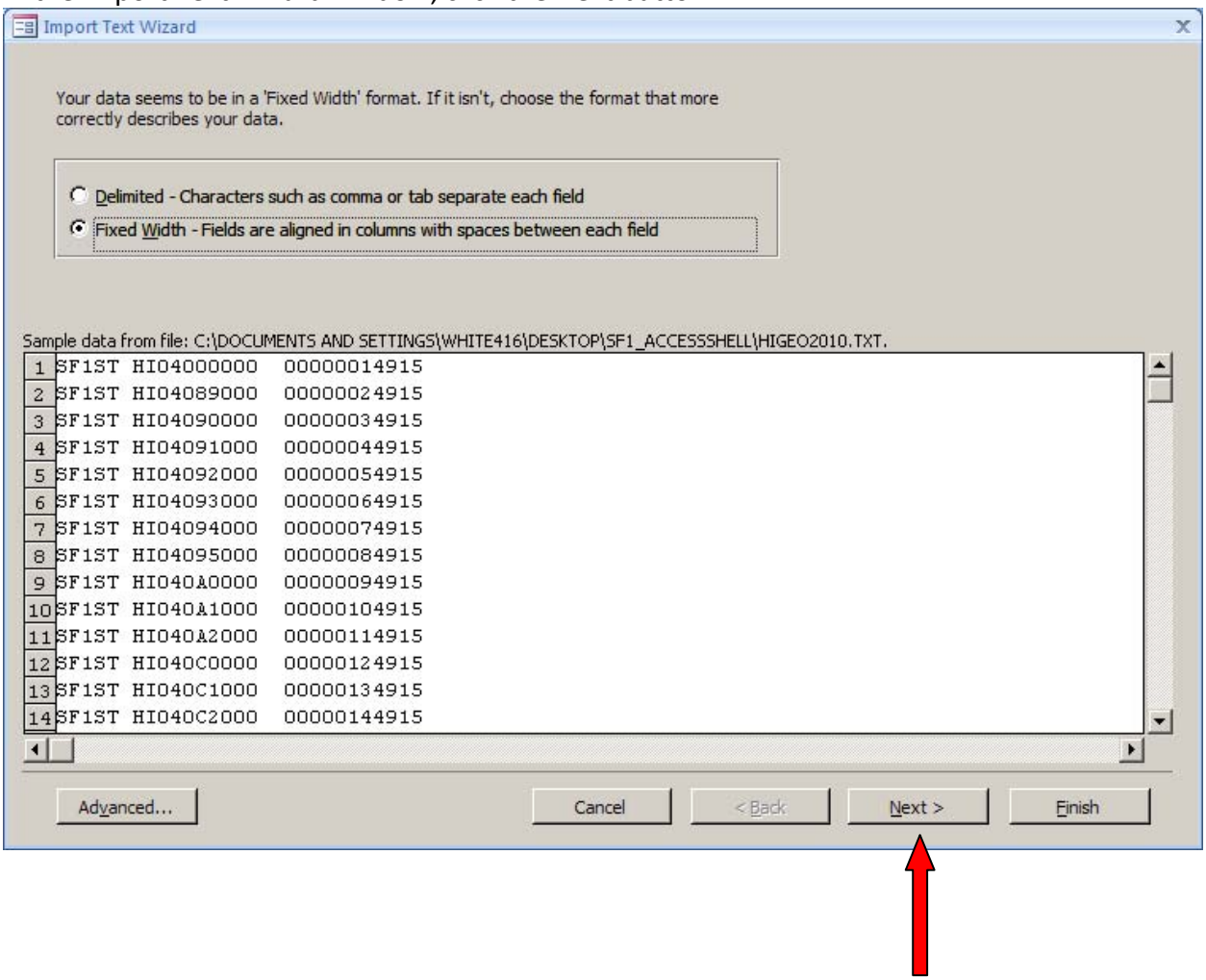
S.) In the Import/Export Specifications window that opens, select the GEO Header Import Specification and click Open.



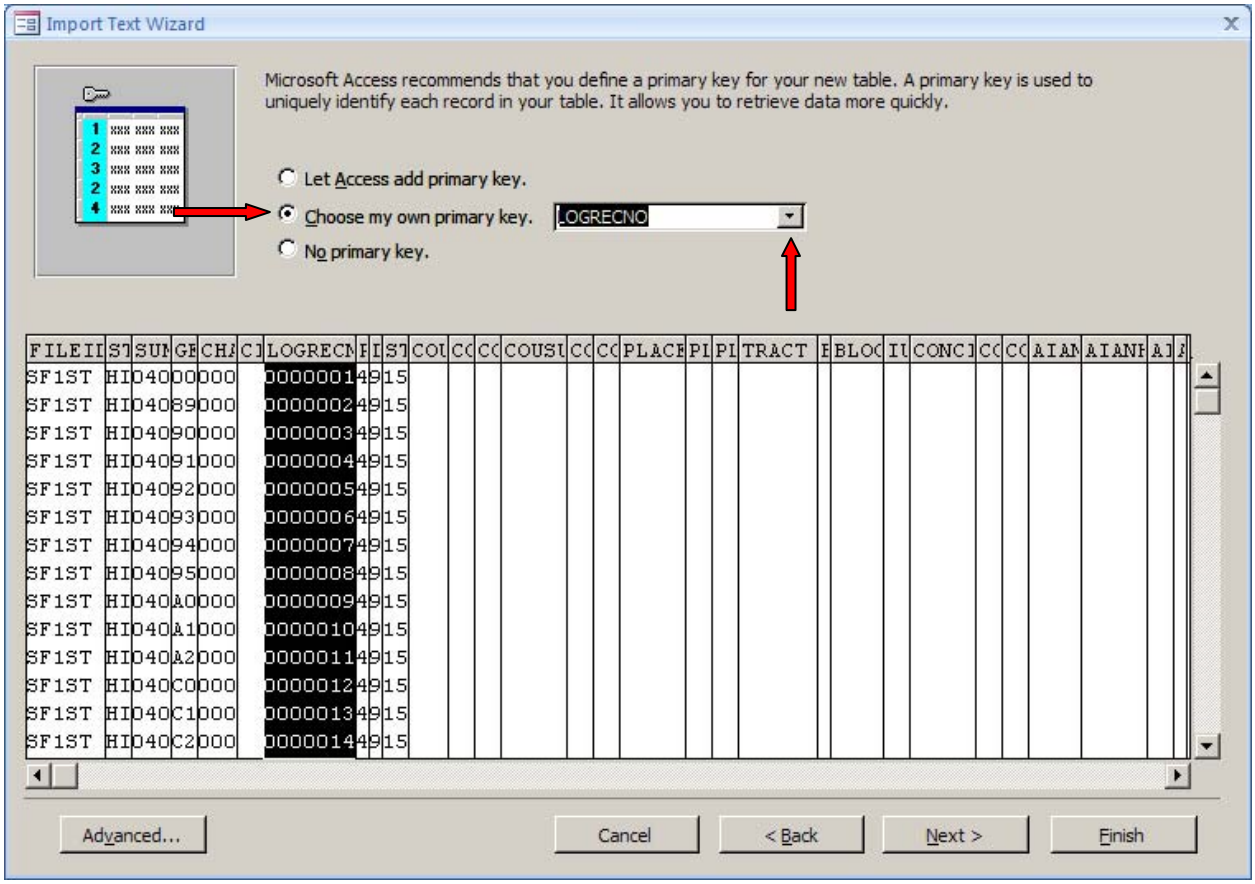
T.) The Import/Export Specification window closes and you are back at the Import Specification window. On the right hand side, click the OK button.



U.) In the Import Text Wizard window, click the Next button.

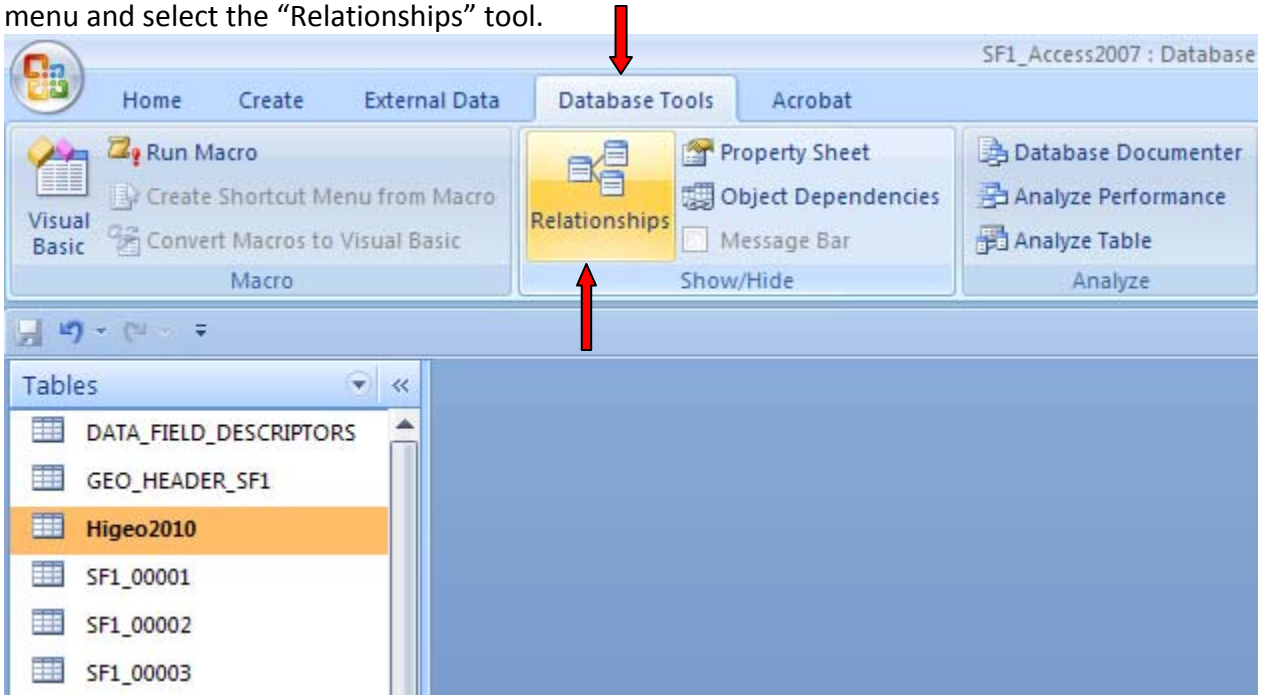


V.) Continue clicking the Next button until you reach the Primary Key setting screen. On this screen, select “Choose my own primary key” and use the pull down arrow to select the LOGRECNO field. Then click the Finish button. You will end up with a geo-header table named <ST>geo2010 where the <ST> is the two letter postal abbreviation for the state.

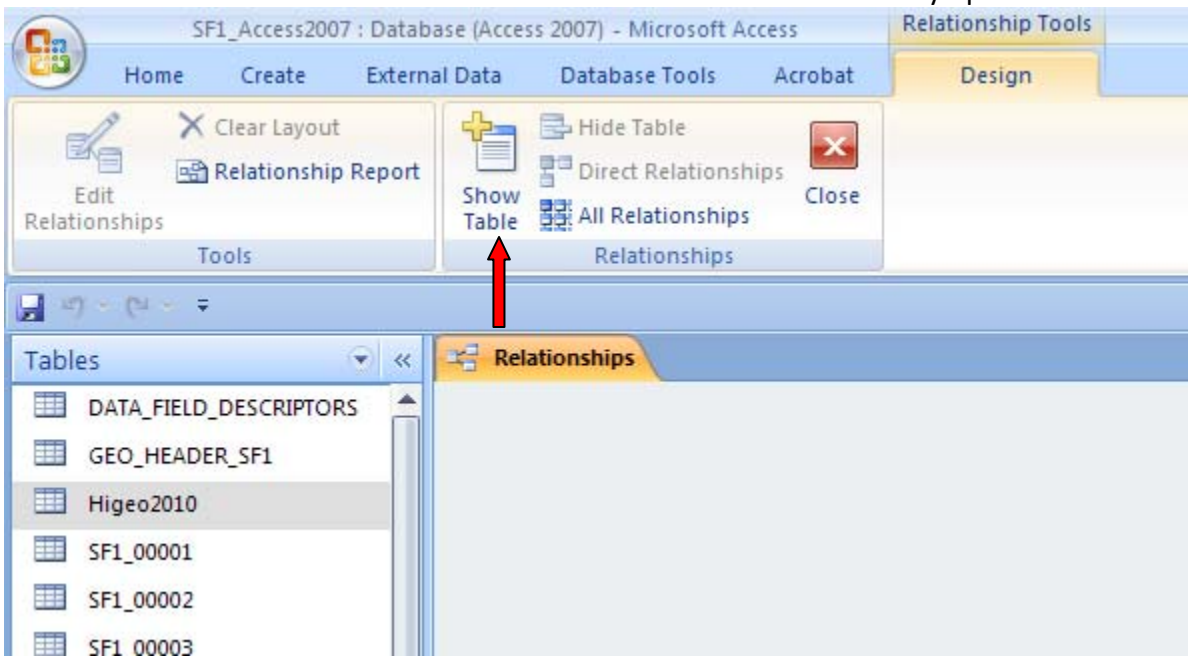


Joining the Data

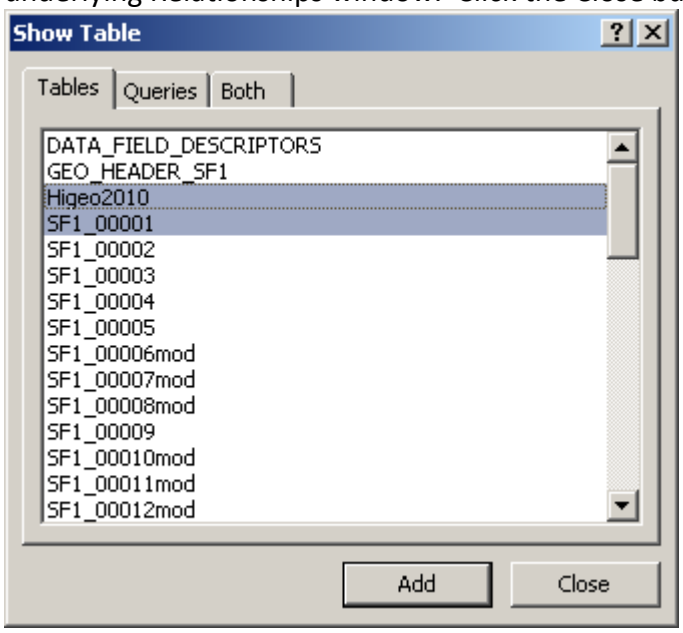
A.) Once the segment and geo-header have been imported, click on the “Database Tools” menu and select the “Relationships” tool.



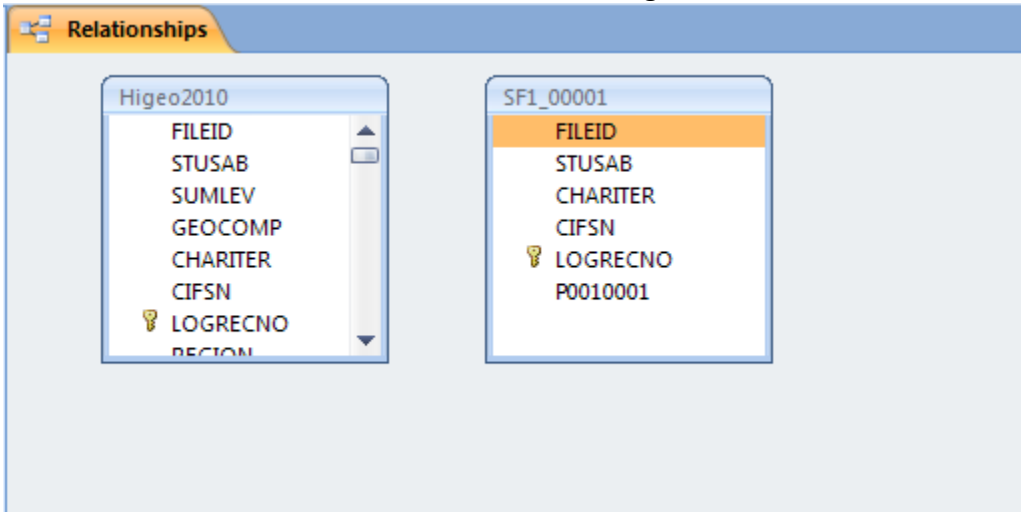
B.) Click on the Show Table button if the Show Table window hasn’t already opened.



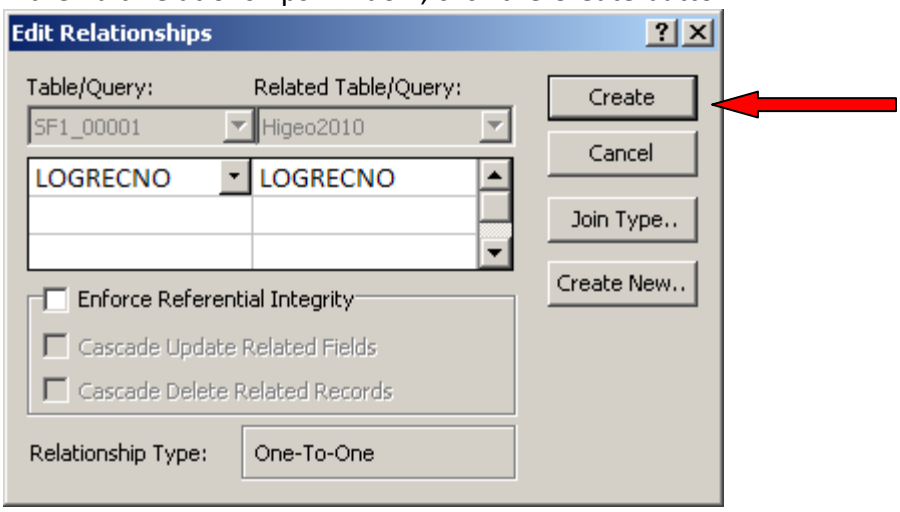
C.) In the Show Table window, highlight the geo header and data segment table you have imported and then click the Add button. Both can be selected at the same time by holding the Ctrl key while clicking on the tables. This will add the tables to the underlying Relationships window. Click the Close button.



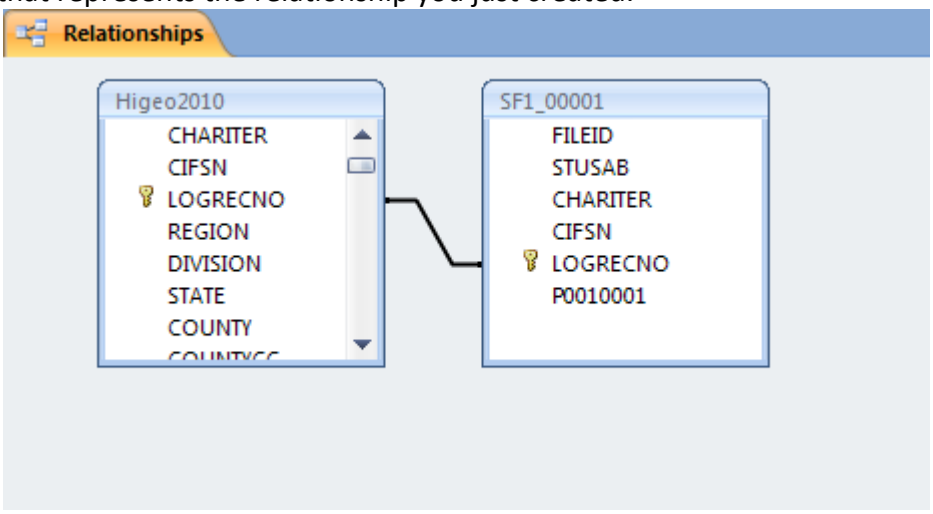
D.) To create a relationship between the tables, it is necessary to link them using the LOGRECNO field. Click and hold on the LOGRECNO field in the Segment table and drag and release that onto the LOGRECNO field of the geo-header file.



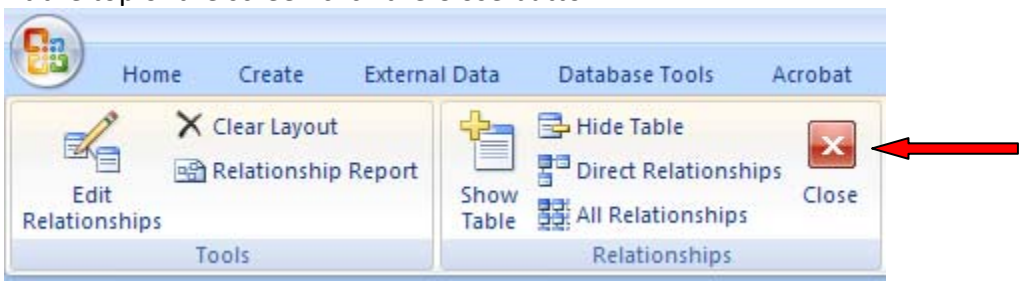
E.) In the Edit Relationships window, click the Create button.



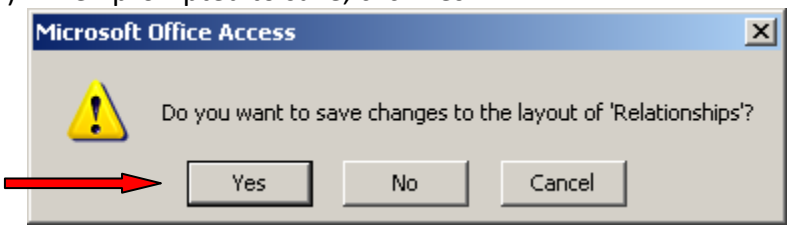
F.) Once completed you will have a line between the Segment table and the geo-header that represents the relationship you just created.



G.) At the top of the screen click the Close button.

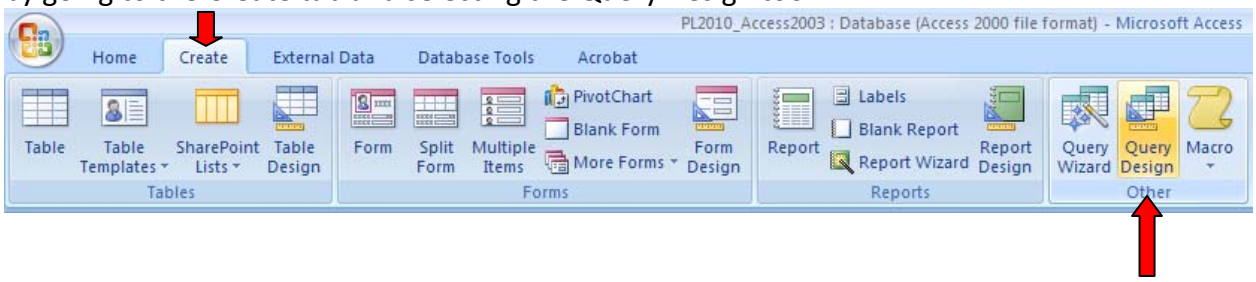


H.) When prompted to save, click Yes.

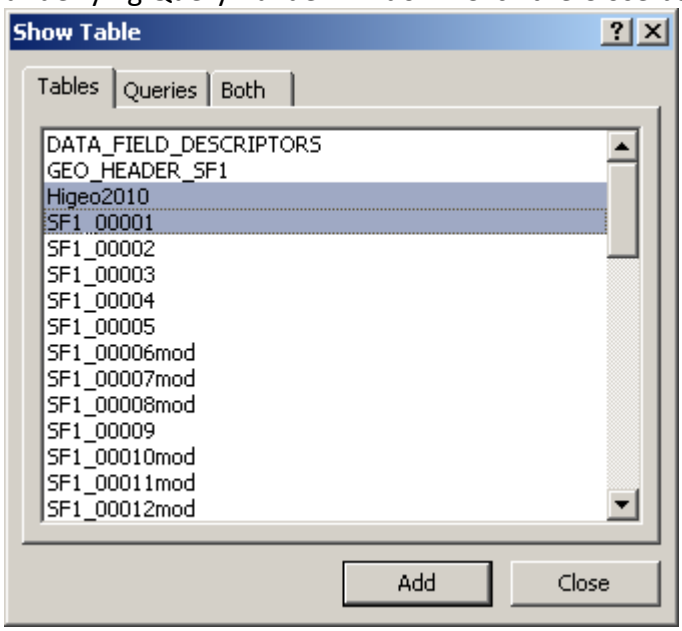


Extracting the Data

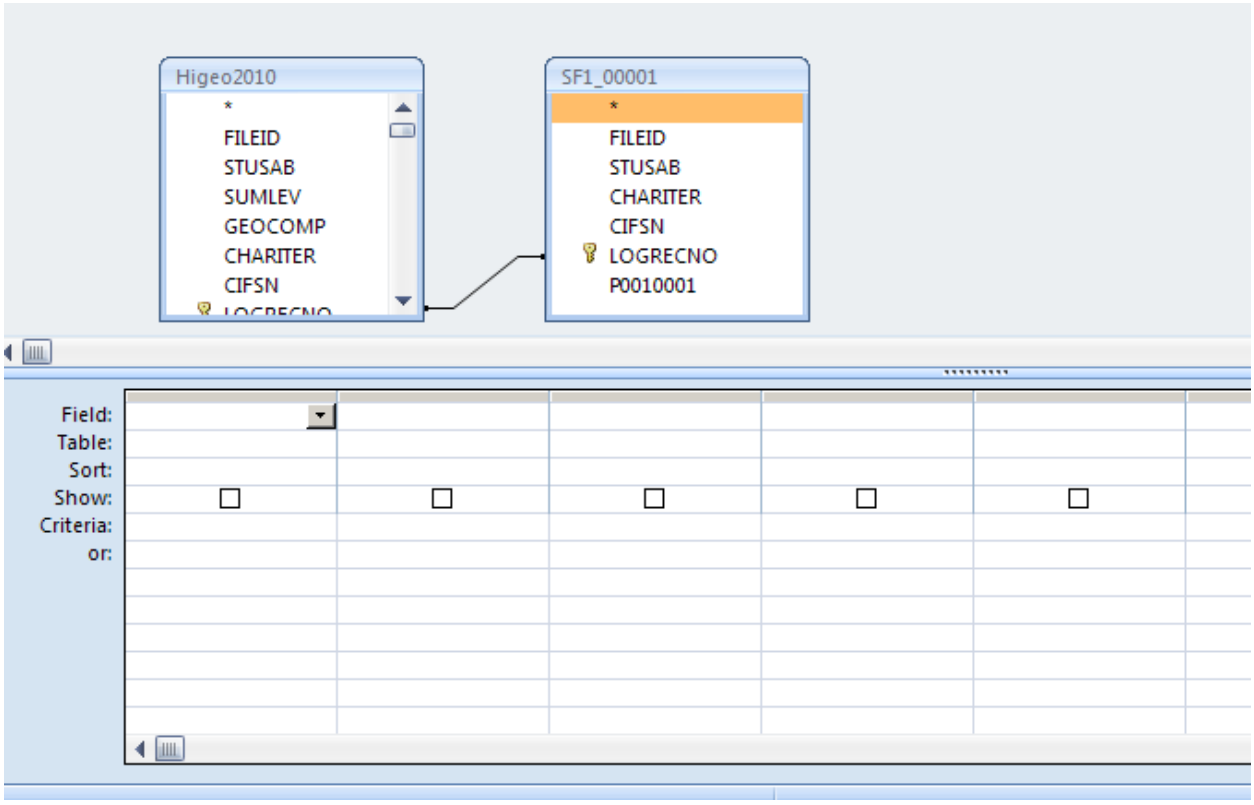
A.) To pull data from the database you just built it is necessary to construct a query. Start by going to the Create tab and selecting the Query Design tool.



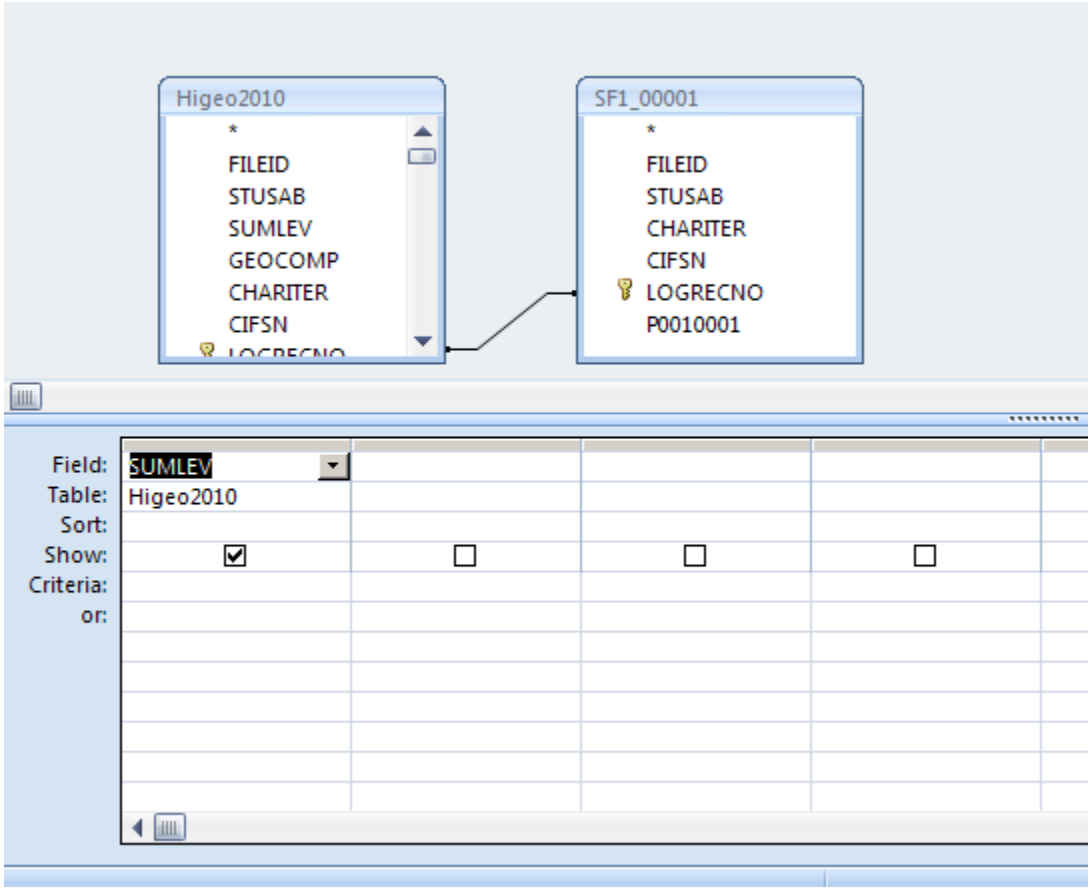
B.) In the Show Table window, highlight the geo header and data segment table you have imported and then click the Add button. Both can be selected at the same time by holding the Ctrl key while clicking on the tables. This will add the tables to the underlying Query Builder window. Click the Close button.



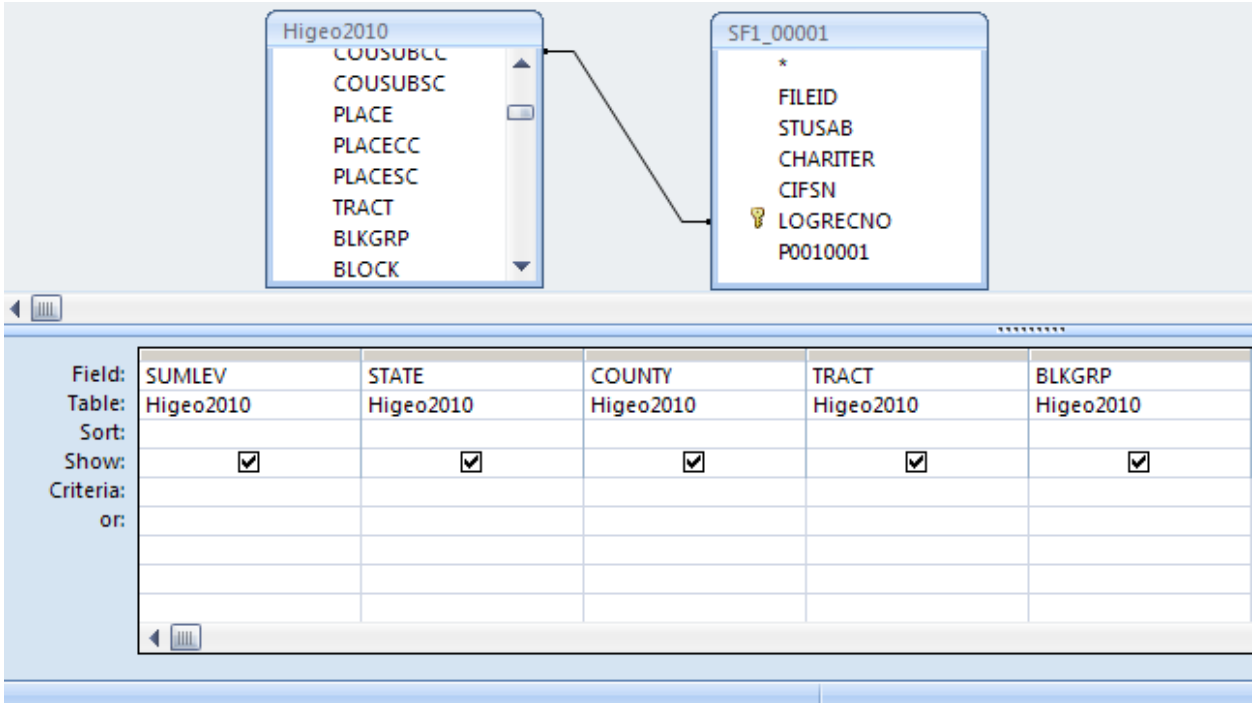
C.) You should now have a window with your two tables displayed in the area above a third table which is shown as a collection of empty columns.



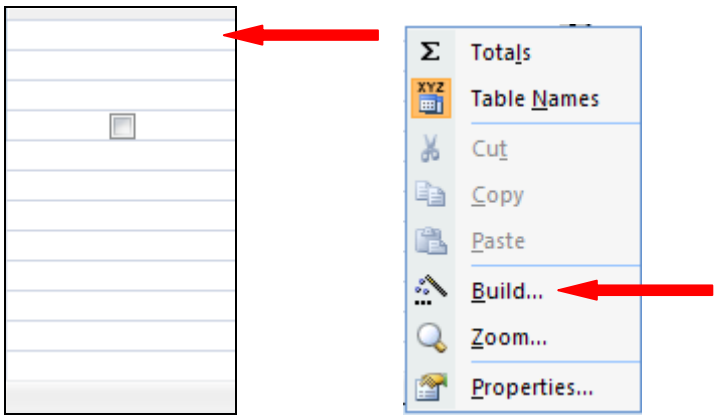
D.) To add fields to the bottom table, double click on the field names in the upper tables. Start building the lower table by adding (double clicking) the field SUMLEV from the geo-header.



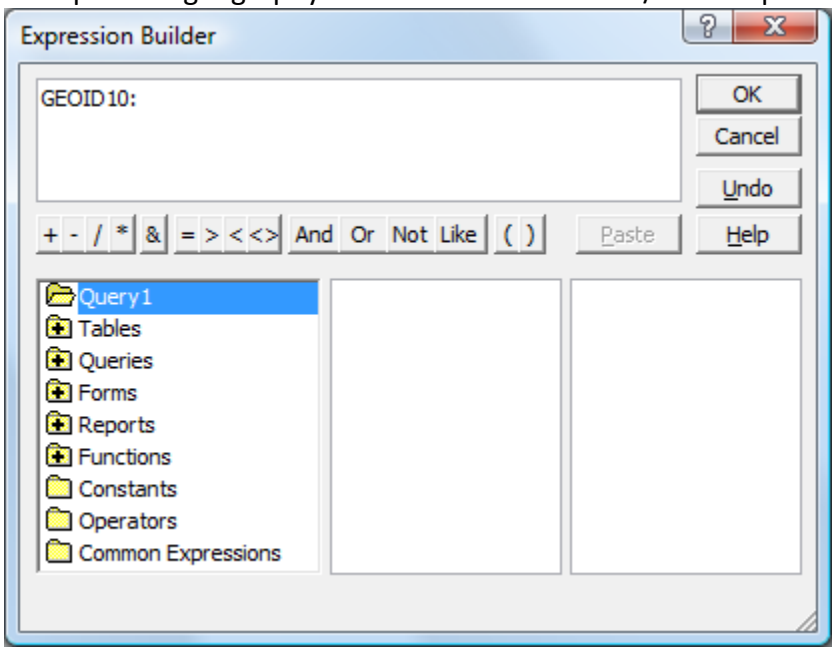
E.) In this example we are building a block group level extract so add all of the fields to the bottom table that go into making a unique GEOID for a block group (STATE, COUNTY, TRACT, and BLKGRP).



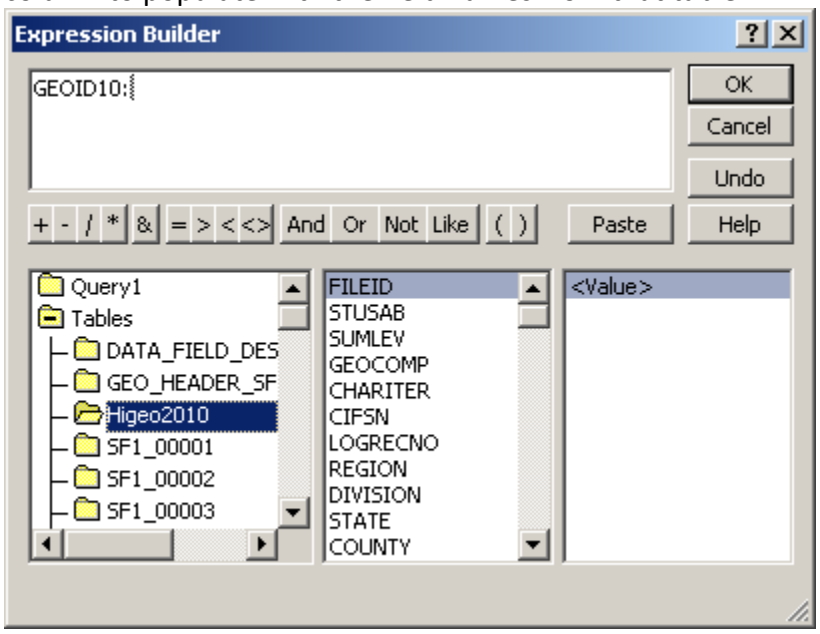
F.) In order to join this data to the TIGER/Line shapefiles, it is also necessary to construct a GEOID. In the bottom table you are building, right click in the first record (Field: row) of the first empty column and select the Build option from the menu that appears. If you are not planning to join this data to the TIGER/Line shapefiles proceed to step K of the Extracting the Data section.



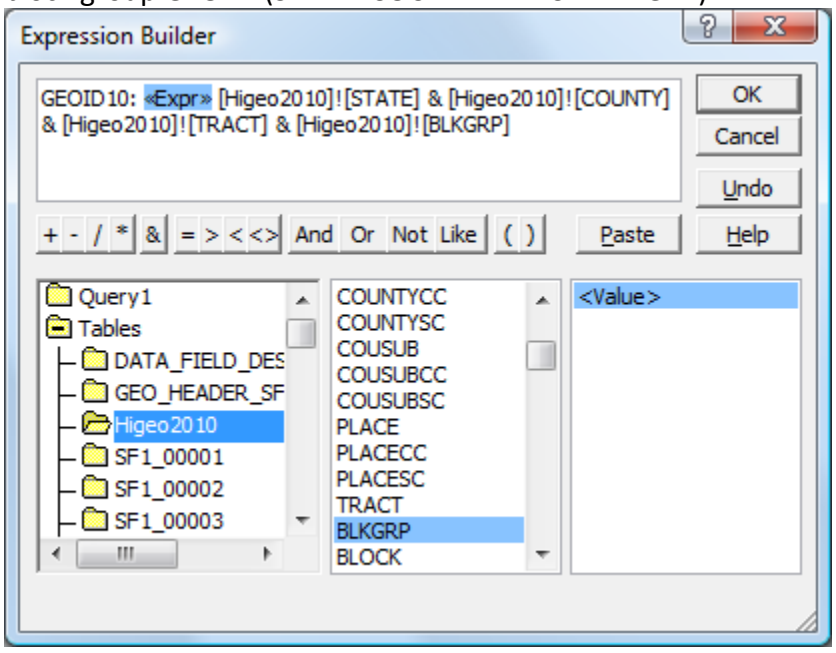
G.) In the Expression Builder type in the name for your GEOID column followed by a colon. This example uses GEOID10. This matches the name of the unique identifier field for each piece of geography found in the 2010 TIGER/Line shapefiles.



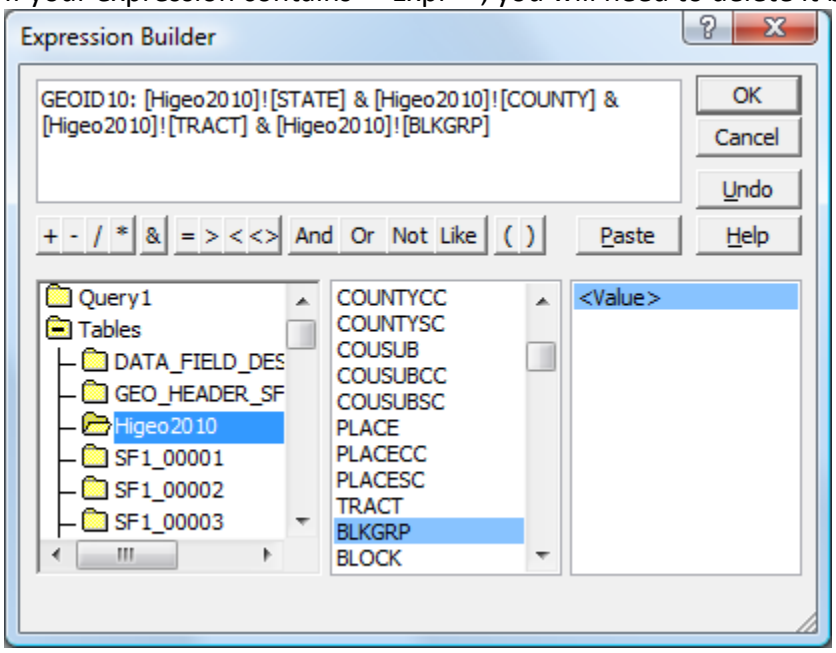
H.) Still in the Expression Builder, in the lower left hand window, double click the plus sign next to Tables and then select the <ss>geo2010 table where <ss> is the two letter postal abbreviation for the state in which you are working. This will cause the middle column to populate with the field names from that table.



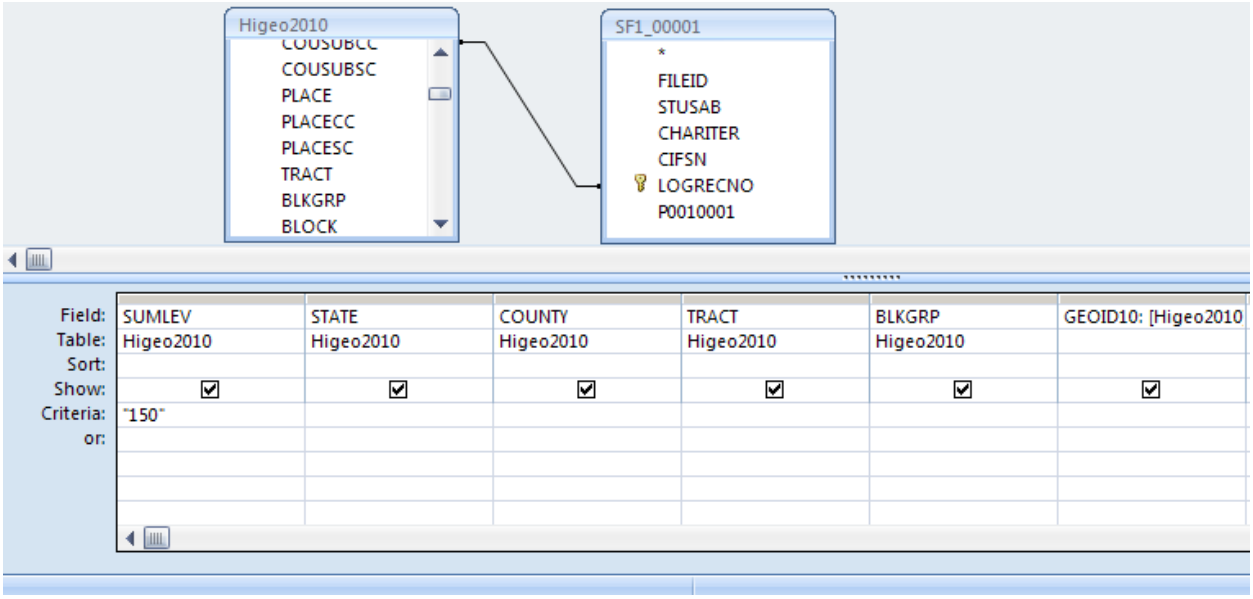
I.) Double clicking a field name in the middle window will move it to the expression window at the top. To concatenate multiple fields, use the ampersand between each field name entry. When you have added all of your fields and ampersands your expression should look similar to the one pictured below. This is an example for a block group GEOID. (STATE+COUNTY+TRACT+BLKGRP)



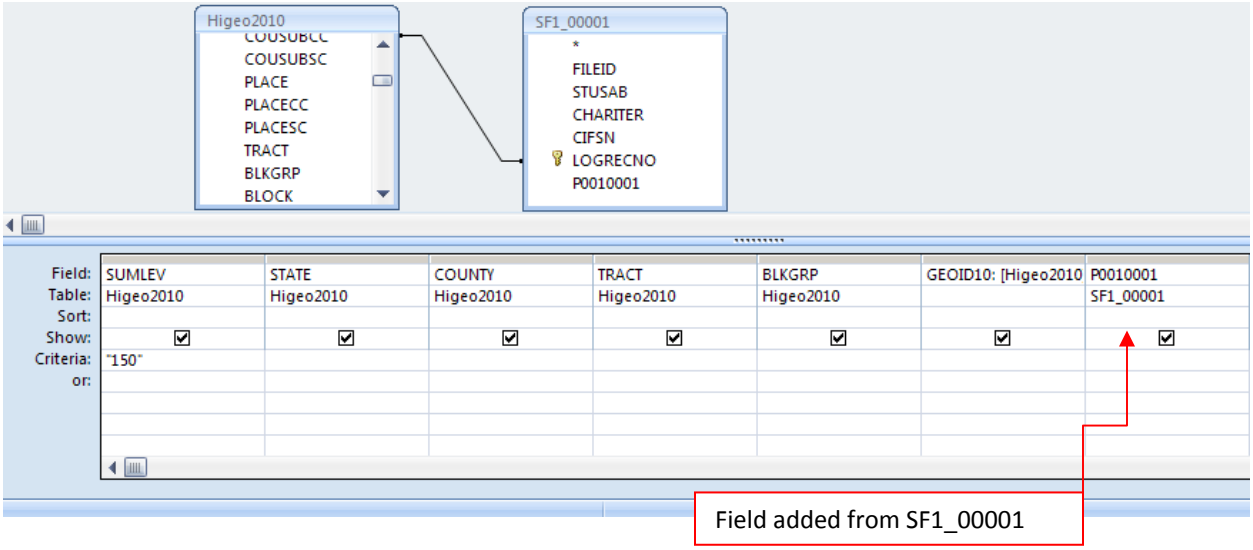
J.) If your expression contains <<Expr>>, you will need to delete it before clicking OK.



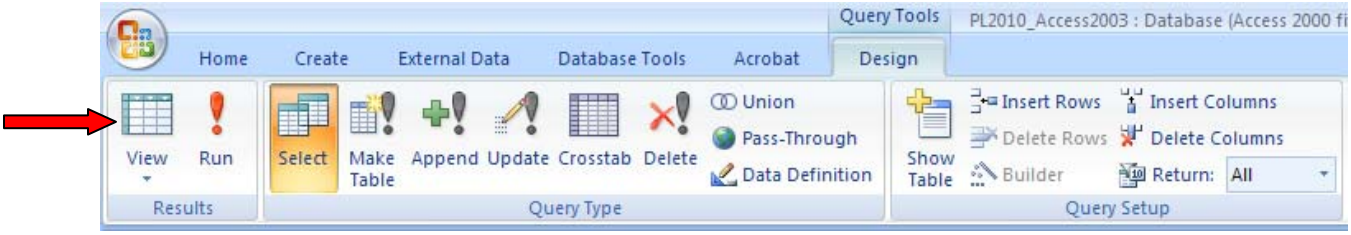
K.) You now have all the geographic components you need for creating a block group level data extract, but now need to restrict the extract to just block group level data. This is done by adding the summary level number for block groups (150) to the SUMLEV field's criteria. Documentation for this information can be found through Item 5 of the Notes and Assumptions section at the beginning of this document. Other fields can be filtered, for example, if you only want one county's block groups you can add that county's three digit FIPS code to the criteria for the COUNTY field.



L.) Now that all the geographic components have been set for the block group level data extract, it is necessary to add the data columns of interest. The data dictionary for the field codes is available in the technical documentation and/or in the Access database shell as a table called "DATA_FIELD_DESCRIPTORS". To add the data columns of your choice, double click them in the upper tables so they appear in the table below.



M.) Once you have added all the data fields of interest, you can view your newly created table. In the upper left hand corner of the Access software window click the View button.

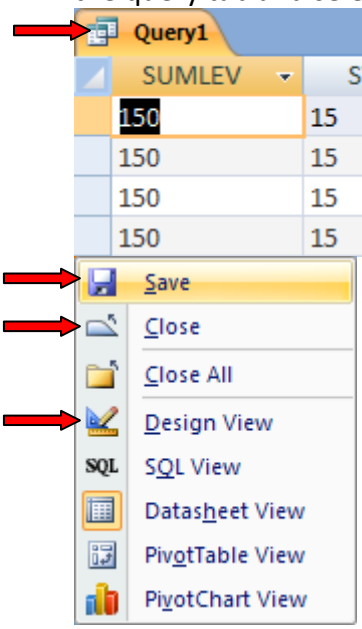


N.) You should now see a table with all geo fields you added, the GEOID10 field you created, and the data fields you added that shows only block group level data records.

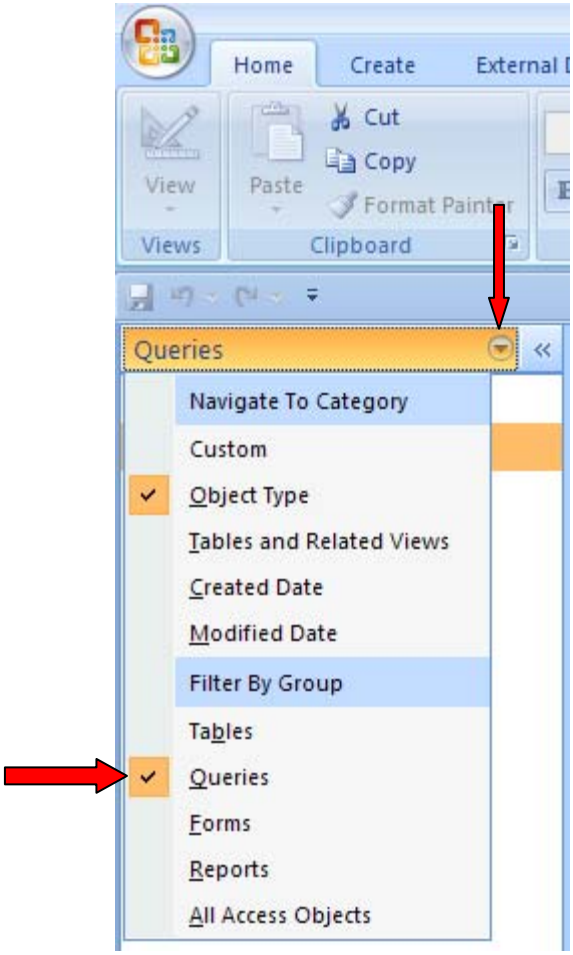
Query1

SUMLEV	STATE	COUNTY	TRACT	BLKGRP	GEOID10	P0010001
150	15	001	020100	1	150010201001	1909
150	15	001	020100	2	150010201002	1010
150	15	001	020100	3	150010201003	925
150	15	001	020100	4	150010201004	1369
150	15	001	020202	1	150010202021	1344
150	15	001	020202	2	150010202022	1224
150	15	001	020300	1	150010203001	992
150	15	001	020300	2	150010203002	1315
150	15	001	020300	3	150010203003	1627
150	15	001	020400	1	150010204001	791
150	15	001	020400	2	150010204002	339

O.) You can now save and close your query for future use by right clicking the Query tab in the upper left of the query screen. To modify an already created query, right click on the query tab and select “Design View”.



P.) If the query is not listed in the Table of Contents on the left hand side, click the banner at the top of the Table of Contents. On the resulting menu pick Queries.



Q.) If you want to export the table you created, right click on name of the query you saved and select Export. This will provide many format options for your exported table. Note: Some formats may not be compatible with the size of the table you are exporting.